Sveučilište u Rijeci – Odjel za informatiku

Preddiplomski studije jednopredmetne informatike

Dorian Beli

Usporedba jezičnih alata za njemački jezik

Završni rad

Mentor: izv. prof. dr. sc. Sanda Martinčić-Ipšić

Rijeka, rujan 2018.

Sadržaj

Predgovor

Veliku zahvalnost prvenstveno iskazujem svojoj mentorici izv. prof. dr.sc Sandi Martinčić-Ipšić koja je u meni probudila zanimanje prema području računalne analize prirodnog jezika, te na ukazanoj pomoći tokom pripreme ovog završnog rada.

Također zahvaljujem se svim prijateljima i dragim osobama koji su iskazali neizmjernu podršku tokom ovih godina, bez kojih ne bih imao toliko nezaboravnih i predivnih trenutaka tokom svog studija.

Isto tako zahvaljujem se Odijelu za informatiku Svaučilišta u Rijeci na stečenom znanju i podršci tokom studija kroz osmijeh i savjete.

Najveću zahvalu posvetio bih svojim roditeljima, koji su bili vječita potpora i zvijzda vodilja u najtežim i najsretnijim trenutcima moga života, te koji su me naučli kako nekročiti malen pod zvijezdama kroz neizmjernu dobrotu, skromnost i bezgraničnu ljubav.

Veliko *HVALA* svima!

Sažetak

Kada govorimo o računalnoj analizi i razumijevanju teksta, alati poput lematizatora, korjenovatelja, obilježivača vrsta riječi te različiti korpusi jezika igraju veliku ulogu u području računalne lingvistike. Alati poput ovih promatraju sintaksu i lingvistiku nekog određenog jezika te što boljom primjenom pravila istih, uz pokoju implementaciju vjerojatnosnih algoritama, nastoje bolje obraditi zadani jezik. U ovom završnom radu obrađujemo 4 najpoznatija korjenovatelja, dva lematizatora te dva obilježivača vrsta riječi njemačkog jezika. Osim teorijske obrade navedenih alata, dotaknut ćemo se i praktične usporedbe navedenih u zasebnom poglavlju na vlastitim tekstovima. Korjenovatelji Snowball, CISTEM, Text::Geramn i UniNE, lematizatori GermaLemma i SMOR te obilježivači TIGER korpusa i Pro3GreDE imaju iskazanu točnost u postotcima. Među korjenovateljima najuspješniji se pokazao CISTEM korjenovatelj s 91.23% točnih korjenovanja, zatim Text::German sa 88,55% kojeg slijedi Snowball sa 82,44% te na kraju UniNE koji ima točnost u rasponu od 78,63% do 80,92%. Između dva obilježivača vrsta riječi točniji se pokazao hibridni Pro3GresDE sa 93,55% te onaj uključen unutar TIGER korpusa sa 90,32% točnosti. Kod lematizatora točnijim se pokazao SMOR sa 94,27% točnosti te nakon njega GermaLemma sa 85,5% točnosti.

Ključne riječi: korjenovatelj, lematizator, obilježivač vrsta riječi, njemački, korpus, računalna analiza njemačkog jezika

Comparative analysis of natural language processing tools for German language

Abstract

While talking about computer analysis and understanding of texts, natural language processing tools like lemmatizers, stemmers, part-of-speech taggers or treebanks of different languages play a great role in the field of computer linguistics. These kinds of tools analyse syntactic and linguistic rules of a specific language and, with the help of these rules in combination with an occasional implementation of probability algorithms, they process the given language. In this thesis I will analyse four most popular stemmers, two lemmatizers and two part-of-speech taggers for the German language. Alongside the theoretical approach of the given tools, we will also test all eight tools on different texts in the dedicated headings. Snowball, CISTEM, Text:German and UniNE stemmers, GermaLema and SMOR lemmatizers and part-of-speech taggers in Pro3GresDE and the ones included in the TIGER corpus have a success rate shown in percentages. Among the stemmers, the most successful one was the CISTEM stemmer with a success rate of 91,23%, followed by Tex:German with a success rate of 88,55%, Snowball stemmer of the German language with a success rate of 82,44% and UniNE stemmer, as the last one, with a fluctuating success rate, between 78,63% and 80,92%. Between the two part-of-speech taggers, Pro3GerDE had the highest success rate of 93,55% which is followed by the part-of-speech tagger included in the TIGER corpus with a success rate of 90,32%. Between lemmatizers, SMOR had the best success rate of 94,27% followed by GermaLemma with a success rate of 85,5%

Key words: stemmer, lemmatizer, part-of-speech tagger, German, corpus, computer analysis of the German language

Eine Vergleichsanalyse zur maschinellen Verarbeitung der natürlichen deutschen Sprache

Zusammenfassung

Wenn man über maschinelle Verarbeitung der Texte spricht, spielen Lemmatisatoren, Stemmer-Algorhitmen, Part-of-speech Taggers und Textkorpus von verschiedenen Sprachen eine große Rolle in Computerlinguistik. Sie bearbeiten Syntax und Linguistik einer bestimmten Sprache und versuchen mit der Verwendung der syntaktischen und linguistischen Regeln in der Kombination mit den Wahrscheinlichkeitsalgorithmen diese bestimmte natürliche Sprache bestmöglich zu bearbeiten. In dieser Bachelorarbeit werden 4 der bekanntesten Stemmer-Algorhitmen, zwei Lemmatisatoren und zwei Part-of-speech Teggers der deutschen Sprache bearbeitet. In bestimmten Abschnitten wird außer der theoretischen Verarbeitung, auch über den praktischen Vergleich zwischen Stemmer-Algorhitmen, Lemmatisatoren und Part-of-speech Taggers mit Anwendung auf eigenen Texten gesprochen. Die Stemmer-Algorhitmen Snowball, CISTEM, Text::Geramn und UniNE, die Lemmatisatoren GermaLemma und SMOR und Part-of-speech Taggers Pro3GresDE und Part-of-speech Tagger im TIGER Textkorpus drücken ihre Genauigkeit in Prozenten aus. Als erfolgreichster zeigte sich der Stemmer-Algorhitmus CISTEM mit einer Genauigkeit von 91,23%, danach Text::German mit einer Genauigkeit von 88,55%, Snowball mit 82,44% und zuletzt der Stemmer-Algorhitmus UniNE, der zwischen 78,63% und 80,92% lag (hängt vom Typ der Lemmatisation ab).

Zwischen den zwei Part-of-speech Taggers zeigte sich Pr3GerDE als genauer mit einer Genauigkeit von 93,55% und danach Part-of-speech Tagger im TIGER Textkorpus mit 90,32% Genauigkeit. Zwischen den Lemmatisatoren war SMOR genauer mit 94,27% Genaugkeit und danach GermaLemma mit einer Genaugkeit von 85,5%.

Schlüsselwörter: Stemmer-Algorhitmus, Lemmatisator, Part-of-speech Tagger, deutsche Sprache, Textkorpus, maschinelle Verarbeitung der natürlichen deutschen Sprache

1 Uvod

Govoreći o računalnoj analizi bilo kojeg prirodnog jezika, potrebni su nam jezični alati kako bismo bili u mogućnosti računalno obraditi tekstualne sadržaje. Pošto računalo kao takvo nema mogućnost samostalnog razumijevanja prirodnog jezika, veliku važnost ovdje igraju jezični alati poput korjenovatelja (eng. *stemmer*), lematizatora, morfoloških analizatora, različiti obilježivači vrsta riječi (eng. *Part-of-Speech tagger*), koji mogu biti uključeni u korištenom korpusu teksta, te parser jezične sintakse. Sve navedene alate uvrštavamo kao osnovne pristupe pri obradi prirodnog jezika (eng. *Natural Language Processing*, skraćeno NLP) (*SAS-Natural Language Processing*).

To je interdisciplinarna grana koja se uvrštava područje umjetne inteligencije, računalne lingvistike te računarstva s ciljem što uspješnijeg računalnog procesiranja prirodnih jezika bez obzira o obliku njihove primjene (tekst, govor, a u posljednje vrijeme i vizualno s pojavom prevođenja natpisa preko kamera pametnih telefona). Prirodan jezik kao takav je iznimno kompleksan, pa kako bismo u ga potpunosti razumjeli potrebno je obuhvatiti različite jezične discipline kao što su morfologija[1], fonetika i fonologija[2], sintaksa[3], semantika[4] i pragmatika[5] te diskurs (govor, razgovor) (Jurafsky, Martin. 2000).

Morfološki analizatori, korjenovatelji, lematizatori, obilježivači vrsta riječi dio su sintaksne analize prirodnog jezika (Jurafsky, Martin. 2000). Gledajući iz aspekta gramatike njemačkog jezika, lematizatori (Jurafsky, Martin. 2000) kao rezultat vraćaju osnovni oblik dane riječi koji ćemo zvati lema, dok korjenovatelji (eng. *stemmer*) kao rezultat vraćaju korijen, koji prema

---

1 Morfologija (grč. *morphe*= oblik i *logos*= riječ) znanstvena disciplina koja se bavi načinima na koje se riječi u nekom jeziku oblikuju i mijenjaju. Najmanja jezična jedinica u ovoj znanstvenoj disciplini jest morfem. (Hentcshel, Vogel: *Deutsche Morphologie*)

2 Fonetika znanstvena disciplina koja se bavi istražuje obilježja ljudskog glasanja. Fonologija lingvistička disciplina koja proučava funkciju glasova u pojedinom jeziku i u jeziku uopće. (*Hrvatska Enciklopedija*)

3 Sintaksa(grč. *syn*= zajedno i *taxis*=uređivanje, *syntaksis*=slaganje u red) jest znanstvena disciplina koja u jezikoslovlju proučava pravila koja upravljaju ustrojem rečenica te određuju njihovu relativnu gramatikalnost (vezano uz termin *gramatika*- pravila koja upravljaju uporabom jezika). (Silić, Pranjković: *Gramatika hrvatskog jezika za gimnazije i visoka učilišta*)

4 Semantika(franc. *semantique*=koji ima značenje) jest lingvistička disciplina koja se bavi opisom značenja u jeziku. (*Hrvatska Enciklopedija*)

5 Pragmatika(grč. *pragma*=djelovati) je znanstvena disciplina koja proučava jezično djelovanje (*Hrvatska Enciklopedija*)

morfološkim pravilima standardne njemačke gramatike odgovara korijenskim morfemima (njem. *Kernmorphem*). Više o ulogama pojedinačnog alata o obradi prirodnog jezika s lingvističkom pozadinom nešto kasnije unutar primjerenih poglavlja (Jurafsky, Martin. 2000).

Cilj ovog završnog rada je praktična upotreba svakog od navedenih alata te usporedba rezultata koje daju na vlastitim primjerima teksta u odnosu na već dane postotke i mjerenja, kako bismo utvrdili točnost djelovanja svakoga od njih, te ih međusobno usporedili.

Rad sadrži teorijski dio, u kojem se opisuje način rada svakog od alata raspoređenih po kategorijama ovisno o pristupu rješavanja problema u poglavljima 2.1 Lematizatori, 2.2 Korjenovatelji i 2.3 Obilježivači vrsta riječi. Također su u poglavljima 3.1 Usporedba obilježivača vrsta riječi, 3.2 Usporedba korjenovatelja i 3.3 Usporedba lematizatora opisani alati s definiranim točnostima i međusobnom usporedbom, dok se rezultati nalaze u privitku pod poglavljima 5.5 Rezultati obilježivača vrsta riječi Pro3GresDE i TIGER korpusa, 5.6 Rezultati korjenovatelja i 5.7 Rezultati lematizatora.

2 Lematizatori, korjenovatelji, obilježivači vrsta riječi – teorija i praktična primjena

2.1 Lematizatori

Lematiztori kao rezultat daju točan korijen iz aspekta lingvistike. Kako bismo se točnije pozabavili lematizatorima kao alatima za obradu prirodnog jezika, potrebno je razumjeti proces lematizacije te o čemu točno govorimo kada kažemo da je neki korijen riječi njemačkog jezika lingvistički pravilan (Jurafsky, Martin. 2000).

Ovdje veliku ulogu igraju načela njemačkog jezika kao takvog. Svaka riječ u njemačkom jeziku ima svoj osnovnu formu. Uzmimo na primjer riječi poput *gefunden, fand, finde* (*Duden, Deutsches Universalwörterbuch*. 2011). Osnovni oblik tih riječi će biti glagol *finden* (nalaziti). Prva forma je njegov particip[6], druga preterit prvog lica jednine, dok je zadnja riječ obilježava prvo lice jednine prezenta. Što to znači za lematizaciju? Kao što vidimo iz priloženog primjera glagola *finden* svi navedeni oblici kao takvi imaju površinski različit izgled (dodavanje prefiksa *ge-* te nastavka *-en* ili *-t* ovisno da li je riječ o jakom ili slabom glagolu, pojavljivanje slova *a* koje nije uključeno u infinitivu tog glagola itd.) dok je njihov osnovni oblik ili lema jednak, što dovodi do dodatnog kompliciranja, jer je ponekad potrebno odlučiti koju ulogu i značenje ima ta riječ u rečenici. Stoga proces lematizacije definiramo kao proces mapiranja oblika riječi u njihove leme (Jurafsky, Martin. 2000).

Iako ćemo na ovaj način dobiti dobre rezultate, to nipošto nije najbolje rješenje uklanjanja svih pogrešaka. Potrebno je promotriti i pragmatičko značenje riječi te ostala područja lingvistike navedena u uvodu. Uzmimo za primjer riječi *übersetzen* (prevesti) i *übersetzen* (presjesti). Na prvi pogled i jedna i druga riječ imaju isti oblik, ali njihovo značenje kao takvo nije nimalo slično. Postoji poveznica između korjenovatelja i lematizatora koju ćemo opisati u nastavku.

6 Particip njemačkog jezika dio je prošlog vremena. On se tvori pomoću pomoćnih glagola *haben* (imati) ili *sein* (biti) + particip perfekta zadanog glagola.

2.1.1 GermaLemma i SMOR lematizator

2.1.1.1 GermaLemma

Ovaj lematizator napravljen je na temelju TIGER korpusa. Markus Konrad iz Berlin Social Science Center-a (njem. *Wissenschaftszentrum Berlin für Sozialforschung*) tvrdi kako je točnost ovog lematizatora otprilike 71%, dok uz korištenje i instalaciju dodatnog paketa uzoraka se postiže točnost do 84% uz uvjet korištenja python verzije 2.7. U svom programu koristio je 90% korpusa kao rječnik lema, dok je ostalih 10% iskoristio kao podatke za testiranje svojeg lematizatora.

2.1.1.2 SMOR lematizator

SMOR lematizator prvenstveno koristi podatke leksikona iz Wiktionary-a koji sadrži 48 000 korijenskih oblika imenica te 5 500 korijenskih oblika glagola (Sennrich, Kunz. 2014). Ovaj lematizator se koristi kako bi se poboljšalo prirodno procesiranje jezika. Kako on djeluje?

Principi lingvističkih modela implementirani unutar SMOR lematizatora su sljedeći:

- Lematizator implementira pristup konkatenacije unutar pravila morfologije, pretpostavljajući kako sufiksi imaju vlastite leksičke ulaze koji diktiraju ograničenje odabira. U njemačkom jeziku afiksi se spajaju s bazom ovisno o vrsti riječi, vrsti korijena te po podrijetlu i kompleksnosti riječi;
- Dodavanje afiksa je primijenjeno spojem konkatenacije i provjeravanja značajki pojedinih riječi. Značajke definiraju svojstva korijena i ograničenja odabira afiksa kada govorimo o dodavanju sufiksa i prefiksa;
- Morfofonološka pravila su primijenjena korištenjem operacije zamijene koja mapira analizu niza znakova u površinske nizove znakova. Također vrijedi i obratno;
- Leksikon igra glavnu ulogu u SMOR lematizatoru. On određuje svojstva korijena u odnosu na vrstu leksičkih ulaza, vrsti riječi, vrsti korijena, podrijetlu, kompleksnosti te klasi infleksije.

SMOR-ova gramatika definira klase vezane uz fleksiju riječi koje pokrivaju većinu njemačkih riječi. Na primjer kada bismo uzeli klasu NNeut_s_e, to bi značilo da govorimo o standardnoj imenici srednjeg roda, koja ima nastavak *-s* u genitivu i nastavak *-e* u množini (Sennrich, Kunz. 2014).

Ono što je ključno naglasiti za ovaj lematizator jest da svoj pristup predviđanju fleksije riječi bazira na više načina (ovisno o implementaciji). Pristup baziran na SLES-u[7], koji se koristi za automatsko preuzimanje leksikona iz korpusa kreiranjem infleksijske hipoteze za svaku riječ korpusa, te odabirom klase infleksije za leksikon temeljen na statistici (Adolphs, 2008).

Prednosti su:

1. Predviđanjem infleksije s podatcima dobivenima iz Wiktionary-a izbjegava se sva nesigurnost vezana uz problem da li dvije riječ imaju istu lemu ili ne;
2. Prepoznaje se gramatička kategorija svake riječi iz tablice infleksije.

S druge strane problem jest taj da se za različite infleksijske klase ne može raspoznati razlika isključivo analizom teksta; ponekad nećemo moći prepoznati rod leme ili kada se nastavak *-s* koristi kao genitivni nastavak, kao nastavak množine ili pak oboje (Sennrich, Kunz. 2014).

Ovaj lematizator funkcionira na sljedeći način:

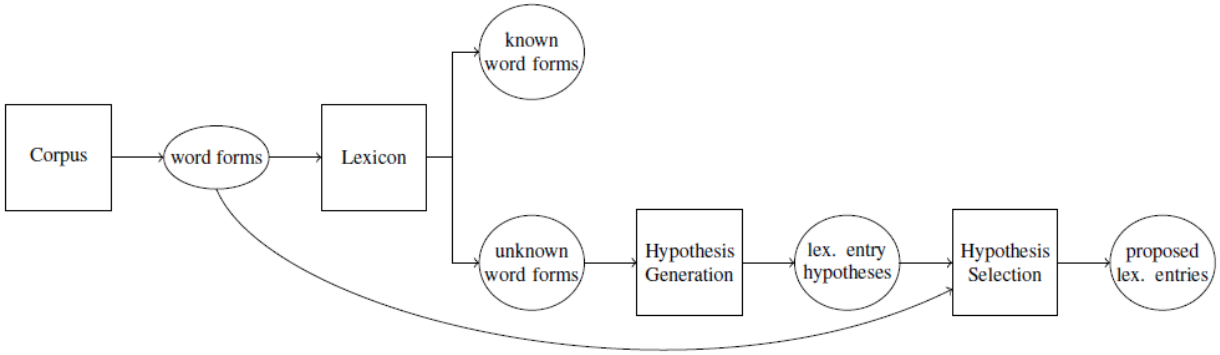
1. korak: Za svaku riječ u tablici fleksije, generira se lista infleksijske klase pomoću SLES-a te se lista filtrira pomoću podataka prikupljenih s Wiktionary-a (npr. za riječ *Hauses* – genitiv riječi *Haus*- SLES vraća 120 hipoteza). Filtriranjem pretpostavki s krivim korijenom, rodom, brojem ili padež smanjuje broj pretpostavki za genitiv jednine na četiri.
2. korak: Pretpostavke za sve riječi u tablici infleksije se presjeku, što daje krajnju klasu infleksije (primjer za riječ *Haus* daje NNeut_es_$er, što znači imenica srednjeg roda sa *-es* nastavkom za genitiv, prijeglasom ili na njem. *Umlaut,* te nastavak *-er* kao nastavak za množinu navedene riječi).

Međutim kako njemački jezik ima i nepravilne glagole, SMOR će u tom slučaju rezultirati praznim setom klase infleksije. U tom slučaju za nepravilne glagole se oslanja na ručno unošenje, dok u slučaju imenica se nastoji predvidjeti infleksija jednine i infleksija množine zasebno te ih nakon toga upisati u leksikon ako oba oblika nisu jednoznačna. Ako je Wiktionary-eva tablica nepotpuna, SMOR može predvidjeti više klasa infleksije, što se često

7 SLES je modul za generiranje hipotetskih fleksijskih klasa za SMOR (Adolphs, 2008)

događa kod imenica u jednini, imenica u množini te pridjeva bez oblika u komparativu, kojima se dodjeljuje posebna klasa infleksije (Sennrich, Kunz. 2014).

Problem kod njemačkog jezika nastaje i homonimima *die See* (more) i *der See* (jezero) ili ako jedna riječ ima dva oblika u množini kao npr. *Lexikon* i njegove množine *Lexiken* i *Lexika*. U tim slučajevima se koriste višestruki ulazi za koje se odvojeno predviđa infleksijska klasa riječi. Pretprocesiranje je bitan korak kako bismo prepoznali različite varijante infleksija, koje su ponekad odvojene zarezom ili pak zagradama. Ako predviđanje infleksije stvara višestruke varijante, tada stvaramo višestruke ulaze u leksikon (npr. *Islam*, koji može, ali i ne mora imati nastavak *-s* u genitivu) (Sennrich, Kunz. 2014).



Slika 1: Shematski prikaz sustava preuzimanja riječi (Adolphs, 2008)

Ako SMOR lematizator analizira oblik riječi derivacijom riječi, kao rezultat daje oblik koji ne odgovara onoj formi koju smatramo lemom riječi (nominativ jednine imenica, infinitiv glagola, itd.). Jedno od rješenja jest odvajanje korijena, kojeg želimo zadržati, i nastavka, kojeg ćemo zamijeniti s normaliziranim oblikom, kroz najduži zajednički podudarni niz originalnog oblika riječi i posljednjeg morfema SMOR analize. Drugo rješenje je spajanje dva pretvornika (eng. transducer): originalnog pretvornika koji je generirao SMOR i pretvornika nastalog procesom derivacije originalnog oblika riječi te se mapira analizirana forma u željenu lemu neke riječi. Derivirani pretvornik je dobiven filtriranjem originalnog pretvornika tako da sadrži isključivo osnovne forme (nominativ jednine, infinitiv ili pozitiv) uklanjajući gramatička obilježja iz analiziranog dijela te invertiranjem pretvornika (Sennrich, Kunz. 2014).

Tako smo recimo za riječ *Ermittlungen* (istraživanja) umjesto:

> Ermittlungen
ermitteln<V>ung<SUFF><+NN><Fem><Acc><Pl>
ermitteln<V>ung<SUFF><+NN><Fem><Dat><Pl>

ermitteln<V>ung<SUFF><+NN><Fem><Gen><Pl>
ermitteln<V>ung<SUFF><+NN><Fem><Nom><Pl>

dobijemo:

Ermittlung<+NN><Fem><Nom><Pl>.

Što se tiče implementacije na temelju SFST pretvornika, baza leži u produljenim regularnim izrazima s varijablama i operatorima za konkatenaciju, konjunkciju, razdvajanje, ponavljanje, kompoziciju, negaciju te zamjenu. Kompajler prevodi specifikacije pretvornika u minimalno završno stanje pretvornika koje se koristi za daljnje analize. Osnovne operacije primjene takvog SMOR-a su konkatenacija morfema, filtriranje sekvenci morfema provjerom značajki riječi, te mapiranju rezultata analize nizova znakova u površinske nizove znakova primjenom fonoloških pravila (Schmid, Fitschen, Heid. 2004).

Kao i u prethodnoj implementaciji, korijeni i prefiksi s drugim afiksima su unutar nekog leksikona. Derivirani oblici se generiraju dodavanjem sufiksa na korijen. Rezultirajući pretvornik je sastavljen od filtera koji provjerava i uklanja sufiks prema uvjetima zadanim značajkama određenog nastavka. Rezultat se označava sa S_0. Derivacija prefiksa se generira dodavanjem prefiksa ispred S_0 i provjerom zadovoljavanja značajki korištenog prefiksa, na taj način rezultirajući pretvornikom P_1. Dodaju se daljnji nastavci na P_1 prema zadanim značajkama tog sufiksa kako bismo dobili S_1. Disjunkcija dobivenih S_0 i S_1 nam daje set jednostavnih i deriviranih oblika (Schmid, Fitschen, Heid. 2004).

2.2 Korjenovatelji

Prethodno smo napomenuli kako djelovanje lematizatora rezultira pravim krojenom riječi te kako, iako uspješni, još uvijek imaju problem određivanja leksičke kategorije riječi, naročito kada govorimo o homografima (primjer iz poglavlja lematizatora *übersetzen*=prevesti i *übersetzen*= presjesti). Kako bismo točnije odredili leksičku prirodu neke riječi te kako bismo mogli pobliže odabrati pravilo normalizacije i uklonili greške, potrebno je primijeniti korjenovatelj. U uvodu sam spomenuo kako korjenovatelji daju „*pseudo-korijen*“ i kako se taj „*pseudo-korijen*“ u njemačkoj gramatici, točnije morfologiji, takav morfem naziva korijenski morfem ili *Kernmorphem* [*Duden, Deutsches Universalwörterbuch*. 2011] na njemačkom jeziku. Prije same definicije korjenovanja i njihova načina rada moramo shvatiti kako u njemačkom jeziku dobivamo korijenske morfeme.

Upoznati smo da tijekom analize riječi neka riječ može imati svoj sufiks i prefiks (na primjeru od glagola *übersetzen*, gdje bi predmetak *-über* bio prefiks, a nastavak *-en* sufiks), međutim njemački ima dvije bitne iznimke kada govorimo o afiksima[8]; infiks i na njemačkom *Zirkumfix*[9]. Infiks prema pravilima njemačke lingvistike generalno ne ovisi korijenu riječi, već je kao takav umetnut između slogova. Promotrimo riječ *einzuschieben* (umetnuti) te ju rastavimo na dijelove. Rezultat ćemo odvojiti okomitom crtom „|“. Rezultat toga bi bio sljedeći: *ein* | *zu* | *shieb* | *en*. Ovakvim razdvajanjem dobili smo da nam je predmetak – *zu* – zapravo infiks ove riječi.

S druge strane imamo i tzv. *Zirkumfixe*. Oni se pojavljuju u dvije situacije njemačke gramatike:

- Derivacije riječi: (primjeri su *Gerenne* (imenica koja simbolizira trčanje uokolo), *Gehetze* (energičnost)) Kao što primjećujemo i jedna i druga riječ, iako imenice pisane velikim početnim slovom, su zapravo korijenski glagoli. Ovdje govorimo o procesu pretvaranja glagola u imenice ili na njemačkom *Substantivierung des Verbs*. U ovom slučaju *Ge-...-e* ima ulogu cirkumfiksa, jer prethodno prikazanom metodom svođenja riječi na morfeme -*renn*- i -*hetz*- su korijenski morfemi.
- Fleksija glagola kroz vremena: primjeri: *gelacht, geehrt, gelaufen, gerungen*; ovdje predmetak *ge*- skupa sa nastavcima *-t* ili *-en* (prvi nastavak za jake, drugi za slabe).

Sada kada razumijemo princip djelovanja njemačke morfologije, otprilike znamo koje korijenske morfeme bi njemački korjenovatelj trebao vratiti kao rezultat. Cijeli ovaj proces uklanjanja sufiksa i prefiksa neke riječi jest proces korjenovanja (eng. *stemming*) (Jurafsky, Martin. 2000).

Točnije svi nastavci neke riječi se jednostavno odrežu od onoga što bi nam trebalo ostati kao korijenska osnova neke riječi. Međutim, postoje situacije u kojima korjenovatelji također griješe, jer cijeli navedeni postupak dobivanja korijena neke riječi (ovdje govorimo o lingvističkom pristupu) nije u potpunosti implementiran u svim korjenovateljima. Neki su samo

---

8 Afiks - nastavak, skupni naziv za prefikse*, sufikse* i infikse*; jezični element koji se dodaje osnovi riječi sa svrhom da joj izmijeni smisao, značenje, funkciju, ulogu, a da ne razbije njezino jedinstvo; često se naziva i *afiksni* (f. affixal) *element*; (Simeon, Rikard, 1969)

9 *Zirkumfix* je diskontinuirani afiks koji se sastoji od dva gramatička morfema koji „obgrle“ korijenski morfem.

orijentirani na prefikse i sufikse pri čemu se infiksi uklanjao posebnim algoritmima korjenovanja, dok nailaze na probleme prijeglasa (njem. *Umlaut*) pri čemu neki glagoli dobivaju jednaki formu. Odgovor na taj problem riješen je uglavnom pomoću pravila za jednostavnu zamjenu i to:

1. zamjena znaka ***ß*** sa ***ss***,
2. zamjena ***ä*** s ***ae***,
3. zamjena ***ö*** sa ***oe***,
4. zamjena ***ü*** sa ***ue***.

Ovo je samo jedan od načina na koje možemo riješiti problem. Ono što je bitno napomenuti jest da većina korjenovatelja njemačkog jezika u potpunosti izbacuje pojmove *Zirkumfix*-a te ih isključivo promatraju sve kao sufikse, što uvelike olakšava proces korjenovanja.

Međutim, korjenovanje njemačkih riječi se također može izvršiti i pogrešno, pa tako imamo situaciju u kojoj dolazi do pretjeranog skraćivanja (skraćivanje dvije različite riječi na jednak korijen, iako tim riječima korijenski morfem nije isti) koje se u nekim situacijama može pojaviti npr. kod osobnih imena (ime *Albrecht Dürer* je skraćeno samo na *Dur*) ili kod zamijene prijeglasa (*Körper* i *Korps* korjenovali na *Korp*) itd., ili premalog skraćivanja (pojava u kojoj se može dogoditi da korjenovatelj nije uklonio cijeli sufiks neke riječi) njemačkih riječi, koji se kod nekih alata pojavljivao kod nepravilnih glagola, riječi čije podrijetlo potječe iz latinskog ili starogrčkog jezika ili čak, što je najuočljivije, kod profesija ženskoga roda (za riječ *Schauspielerinnen* pokazuje *Schauspielerinn* kao korijen riječi, što je krivo). Neki od ovih problema su riješeni na vrlo jednostavan način zamjene, što je u konačnici potpomoglo dobivanju boljih rezultata (Caumanns. 2001).

Dakako postoje različiti pristupi korjenovanju riječi, pa tako algoritme korjenovanja može podijeliti na stohastičke, hibridne, te stemere skraćivanja. Svakoga od njih ćemo ukratko opisati u zasebnim poglavljima (Jurafsky, Martin. 2000).

2.2.1 Stohastički algoritmi

Ovakvi tipovi algoritama koriste statističke modele kako bi se prepoznao korijen neke riječi. Ovdje govorimo o algoritmima koji se „uče“ na temelju tablice riječi u kojoj se nalazi njen korijenski oblik u odnosu na njen originialni oblik na koju bi korjenovatelj trebao utjecati kako bi se stvorio vjerojatnosni model. Riječ o modelu koji je u napravljen u skladu s lingvističkim pravilima, točnije, govorimo ili o uklanjanju afiksa iz riječi kako bismo dobili korijenski oblik

ili o procesu lematizacije same riječi, pri čemu se korjenovanje stemerom u potpunosti izbjegava ili se pak radi o primjeni dva pravila sekvencijalnim putem, pri čemu se odluka o korištenju određenog postupka bazira isključivo na vjerojatnosti koliko će rezultat svake navedene opcije rezultirati točnijim rješenjem našega početnog problema (Jurafsky, Martin. 2000).

Najpoznatiji predstavnici ove skupine su tzv n-grami koji mogu biti fonemi, slova, slogovi, pa čak i cijele riječi ili minimalni parovi riječi nekog jezika, sve ovisno o njihovoj primjeni (Jivani, 2016). Nazivima n-grama dodajemo latinske numeričke prefikse, pa tako n-grami kod kojih je n jednak 2 zovemo bigrami, n-grami kod kojih je n jednak 3 trigrami itd. Na koji način oni funkcioniraju? Pretpostavimo da su nam n-grami slova, te da svakom od tih slova pridružimo određenu vjerojatnost koja nam govori kolika je mogućnost da će se baš to slovo pojaviti sljedeće. Sveukupan zbroj vjerojatnosti tih slova je jednak 1. Što znači ukoliko bismo gledali ortografske bigrame njemačke riječi *Schemtterling* bi izgledale ovako:

S, SC, CH, HM, ME. ET, TT, TE, ER, RL, LI, IN, NG, G

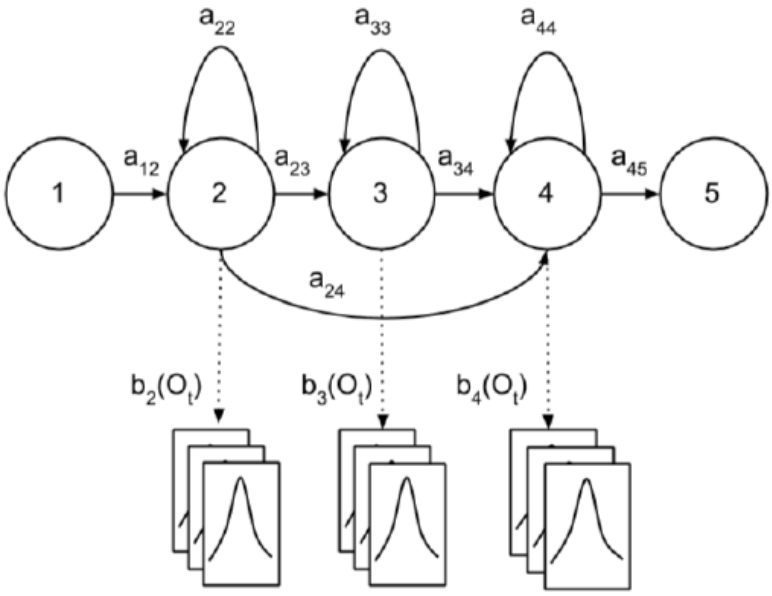
ili ukoliko bi govorili o ortografskim trigramima iste riječi tada bi oni izgledali ovako: **S,*SC, SCH, CHE, HEM, EMT, MTT, TTE, TER, ERL, RLI, LIN, ING, NG* G**,

pri čemu nam zvjezdica označava slobodno mjesto (Jivani, 2016).

Ovo je zasad bio trivijalan problem, međutim što se događa kada govorimo o malo kompliciranijim složenicama njemačkog jezika koje nisu samo spojnica dvaju riječi koje su isključivo osnovni oblici, tj. pravi korijeni u međusobnoj kombinaciji? Uzmemo li za primjer riječ *Anfangspunkt* (na njem.početna pozicija). Standardnim korjenovanjem stemer će ovu riječ razbiti na dvije i to *anfangs* (početak) i *Punkt* (točka) što je krivo razdvajanje tih riječi. Ponekad kod tvorbe složenica kao *Anfangspunkt* ili *Ausbildungszeit* primjećujemo ubacivanje sufiksa ***-s*** radi lakšeg izgovora pojedinih složenica njemačkog jezika. Kako bismo prepravili ovu pogrešku, potrebno je zasebno definirati pravilo koje će nam u tim situacijama reći da nam složenica ima formu ($w_1 + s + w_2$) pri čemu su nam w_1 i w_2 definirane kao riječi koje su nam pohranjene u rječnik. Tek tada, ako je raščlamba naše složenice bila točna, riječ (w_1s) možemo dodati u naš rječnik. Tim putem smanjujemo količinu OOV (skraćeno od eng. *Out-of-*

vocabuary words)[10] te je samim time pogreška WER-a (skraćeno od eng. *Word error rates*) znatno smanjena (Hecht, Riedler, Backfried, 2002).

Svakako kada govorimo o stohastičkim algoritmima n-grammi definitivno nisu jedina mogućnost implementacije. Druga opcija su tzv. skriveni Markovljevi modeli (HMM, eng. *Hidden Markov Models*)[11] generalno. Kada govorimo o HMM-u, govorimo o statističkom Markovljevom modelu, u kojem se sustav modelira pod pretpostavkom da je riječ o Markovljevim procesima[12] s nepromatrenim (skrivenim) stanjima (Wang, 2015).



Slika 2: Primjer skrivenog Markovljevog modela sa 5 stanja (Wang. 2015)

Na slici 1 vidimo trivijalan primjer HMM-a s 5 stanja. Stanja su na slici predstavljena s krugovima, te ćemo ih predstaviti sa slučajnom varijablom $S_t \in \{1, 2, 3, 4, 5\}$. Također

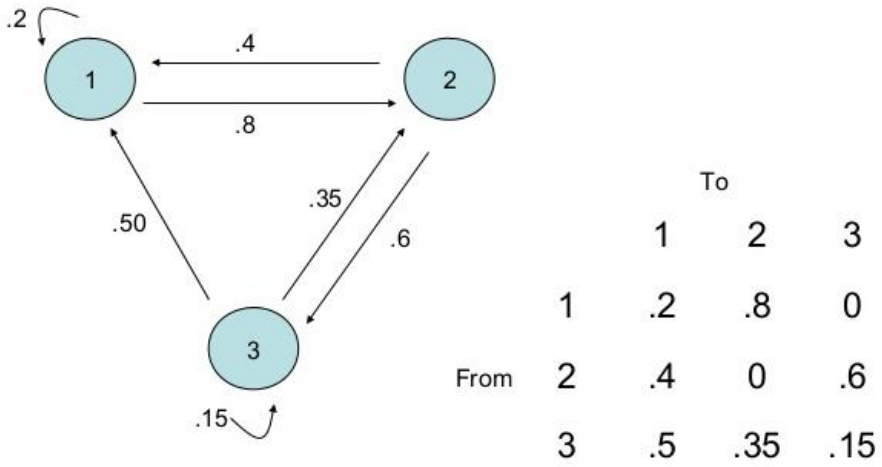
10 Ovdje govorimo o riječima koje originalno nisu bile uključene u našu bazu tijekom stvaranja našeg rječnika.

11 Markovljev model je stohastički model koji se koristi za modeliranje sustava koji vrše slučajnu izmjenu stanja. Pretpostavlja se da buduće stanje ovisi isključivo o trenutnom stanju, a ne o stanjima koja su prethodila trenutnom (Gagniuc, 2017).

12 Najjednostavniji Markovjev model koji modelira stanje sustava sa slučajnom varijablom koja se mijenja kroz vrijeme (Gagniuc, 2017). Također definiran kao stohastički model koji opisuje niz mogućih doađaja u kojima vjerojatnost svakog događaja ovisi o isključivo o stanju koje je postignuto u prethodnom događaju (*Oxford Dictionaries* 2017).

definiramo nasumičnu varijablu O_t koja će nam biti promatranje u nekom vremenu t koje je generiralo trenutno stanje s vjerojatnošću $b_j(O_t)$ (na našoj slici govorimo o $b_2(O_t)$, $b_3(O_t)$ i $b_4(O_t)$ koji korespondiraju sa stanjima 2, 3 i 4). Budući da nam promatranje O_t ne može reći koje djelovanje stanja je rezultiralo njegovim nastankom, praktički dobivamo razlog zašto u nazivu ovog modela imamo riječ „skriveni“. Strjelice u ovom modelu predstavljaju vjerojatnost prijelaza nekih stanja, što znači da se isključivo može doći u stanje iz onog stanja koje ima strjelicu koja upućuje na njega. Varijabla a_{ij} predstavlja vjerojatnost prijelaza iz stanja i u stanje j. Uvjetna raspodjela vjerojatnosti skrivene varijable S_t +1 u vremenu t +1 (buduće stanje) ovisi samo o vrijednosti skrivene varijable S_t u vremenu t (trenutno stanje), što znači da su nam vrijednosti prije vremena t irelevantne (Markovljeva pretpostavka). Slično tome vrijednost promatrane varijable O_t ovisi samo o vrijednosti skrivene varijable S_t, i na vrijeme t. (Wang, 2015)

Kada smo ovo definirali potrebno je nadalje definirati matricu vjerojatnosti prijelaza $N \times N$ (Markovljeva matrica[13]) kod koje i-ti redak matrica zadovoljava uvjet $\sum_{j=1}^{N} \boldsymbol{a_{ij}} = \mathbf{1}$, pri čemu vrijedi $1 \leq i \leq N$ (Wang, 2015).



Slika 3: Primjer Markovljeve matrice (preuzeto sa https://www.slideshare.net/ganith2k13/markov-analysis)

13 Vjerojatnosna matrica koja se koristi za opis prijelaza Markovljevih procesa

Kada smo definirali matricu prijelaza, nadalje je potrebno definirati vjerojatnosni vektor koji može predstavljati vjerojatnost pronalaska sustava svakog stanja. Ulazi tih vektora ne mogu biti negativni te je zbroj njihovih vjerojatnosti mora biti 1. (HaraGopal, 2013)

Ovakav pristup ne zahtijeva lingvističku pozadinu. Kako bismo primijenili metodu skrivenih Markovljevih modela u korjenovanju riječi, potrebno je niz slova koji tvori neku riječ njemačkog jezika definirati kao rezultat povezivanja dvaju podnizova, a to su prefiks i sufiks. Način modeliranja ovog procesa kroz HMM gdje su stanja podijeljena u dva odvojena seta. Prvi set nam mogu biti samo korijeni, dok drugi set mogu biti korijeni ili sufiksi. Prijelazi između stanja definiraju proces izgradnje riječi. Ovdje možemo postaviti određene pretpostavke:

1. Početna stanja pripadaju isključivo setu stemova, što znači da riječ uvijek počinje sa stemom.
2. Prijelazi iz stanja koje obuhvaća set sufiksa u stanje seta stemova uvijek imaju null vjerojatnost, što znači da neka riječ može biti samo spoj korijena i sufiksa.
3. Završna stanja pripadaju u oba seta – stem može imati niz različitih izvoda, ali također postoji situacija u kojoj stem (korijen) ne sadrži nikakav sufiks.

Za bilo koju riječ, najvjerojatniji put od početnih do završnih stanja će proizvesti točku razdvajanja (prijelaz iz korijena u sufikse), što znači da se dio prije točke razdvajanja smatra krojenom. Nažalost ova metoda može dovesti do pretjeranog skraćivanja, te je kompleksna za primjenu (Jivani 2016).

2.2.2 Algoritmi skraćivanja

Algoritmi skraćivanja, za razliku od prethodnih algoritama ne zasnivaju se na vjerojatnosti, već su bazirani na principu uklanjanja afiksa neke riječi kako bismo dobili korijen neke riječi. Jedan od poznatiji algoritama ove vrste jest Porterov korjenovatelj. Što se tiče verzije Porterovog korjenovatelja za njemački jezik, većina funkcija su preuzete iz onog za engleski jezik, s time da razlika leži u tome da se Porterov korjenovatelj njemačkog jezika ograničava isključivo na uklanjanje sufiksa neke riječi (Weißweiler, Fraser, 2017).

Bitno je naglasiti kako je većina korjenovatelja su usko povezana s Porterovim korjenovateljem i njegovim načinom rada. Svi koriste pravila Porterovog korjenovatelja engleskog jezika uz par manjih promjena vezanih za prijeglase njemačkog jezika te druge iznimke. Većina korjenovatelja ima jednak pristup proceduri, ponekad koristeći gotove algoritme (najčešće

Snowball korjenovatelja) zbog zbog generalnog pristupa korjenovanju riječi njemačkog jezika. Ovdje govorimo o pet korjenovatelja njemačkog jezika: Text::German korjenovatelj[Jivani. 2016], CISTEM korjenovatelj [Weißweiler, Fraser. 2017], Snowball korjenovatelj[*Snowball Tartarus*], UniNE korjenovatelj [Savoy. 2000], te jedan od najznačajnijih, Caumannsov korjenovatelj [Caumanns. 1999].

Princip rada je sljedeći; definiraju se dva područja djelovanja R1 i R2, pri čemu R1 predstavlja područje djelovanja nakon prvog znaka koji nije samoglasnik, koji slijedi nakon samoglasnika ili je null područje djelovanja na kraju riječi ako nema takvog slova koje nije samoglasnik, dok je R2 definiran na isti način s razlikom da se primjena od R2 definira unutar R1. Nakon definiranja R1 i R2 briše se određeni broj sufiksa ako se oni pojavljuju unutar R1 ili R2. Ovaj postupak ne vrši se rekurzivno, već u tri koraka u svakom od kojih se može ukloniti najviše jedan sufiks. Prva dva koraka uklanjaju česte sufikse kao što su *-ern* ili *-est*, dok treći korak uklanja sufikse koji se rjeđe koriste kao što su ti sufiksi *-isch* ili *-keit*. (Weißweiler, Fraser, 2017)

2.2.2.1 Caumannsov korjenovatelj

Kako bismo razumjeli na koji način funkcionira ovaj korjenovatelj važno je napomenuti kakve prepreke je potrebno prijeći kako bismo uspješno korjenovali neku njemačku riječ. Tu ćemo se dakako ponovno dotaknuti područja njemačke lingvistike.

Konkretno, ovdje postoji par problema s lingvističke strane koje prvenstveno treba uočiti kada govorimo, ne samo o množini imenica njemačkog jezika, nego i generalno o nastavcima koji se pojavljuju u njemačkom jeziku (*Duden, Deutsches Universalwörterbuch*. 2011). Ti problemi su sljedeći:

1. Promjene u korijenu se ponekad ne događaju na samom njegovom kraju, nego ponekad i u sredini (pravi primjer za to je množina riječi *Haus* = kuća koja glasi *Häuser*, prijelaz iz ***a*** u ***ä*** uz nastavak ***-er***)
2. Sva pravila deklaracije imenica su bazirana na rodu. Bez dublje leksičke analize ili rječnika nemoguće je utvrditi da li je npr. sufiks *-er* (kao što je nastavak u riječi *Bilder*) ili je pak dio korijena (kao u riječi *Leber*).
3. Postoje mnoge iznimke kada govorimo o zamijeni samoglasnika za prijeglas (njem. *Umlaut*) pomoću kojeg se tvori množina (npr. *Hund* u množini glasi *Hunde*, ali *Mund* u množini glasi *Münder*).

4. Zadnji od problema su složenice njemačkog jezika. Za primjer možemo uzeti najdužu njemačku riječ koja dolazi iz područja prava vezano za regulaciju testiranja govedine, koja ima 63 slova, a glasi *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz.* Riječ je 2013 izbačena iz korištenja zbog njezinog predugog naziva (Caumanns, 2008).

U njemačkom jeziku imamo sedam sufiksa deklinacije: *-s, -es, -e, -en, -n, -eri* i *-ern*, 16 za pridjeve: ***-e***, ***-er***, ***-en***, ***-em***, ***-ere***, ***-erer***, ***-eren***, ***-erem***, ***-ste***, ***-ster***, ***-sten***, i ***-stem***, dok ih za glagole ima čak 48: ***-e***, ***-est***, ***-st***, ***-et***, ***-t***, ***-en***, ***-ete***, ***-te***, ***-etest***, ***-test***, ***-eten***, ***-ten***, ***-etet***, ***tet***, ***-end-***, i ***-nd-***, (***-end-***, i ***-nd-*** pretvaraju glagole u priloge te mogu biti popraćeni s bilo kojim sufiksom pridjeva). Predmetak ***-ge*** se također uklanja bez obzira igrao li on ovdje ulogu sufiksa ili prefiksa. Potrebno je uzeti u obzir da imamo još 4 ograničenja koja uzimaju u obzir duljinu i o kojoj situaciji o pojmu govorimo:

1. Ako je pojam kraći od četiri slova ne uklanja se više niti jedno slovo,
2. Ako je pojam kraći od pet slova ne uklanjaju se niti nastavak ***-em*** niti nastavak ***-er***,
3. Ako je pojam kraći od pet slova ne uklanja se nastavak ***-nd***,
4. Nastavak ***-t*** se ne uklanja iz pojma koji započinje velikim početnim slovom. Razlog jest taj da ***-t*** je sufiks glagola te bi njegovo uklanjanje kod imenica njemačkog jezika (jer imenice započinju velikim početnim slovom) ne pridonosi učinkovitosti.

Ovim ograničenjima ne obuhvaćamo riječi sastavljene od dva slova pa će algoritam ignorirati njihovu pojavu. Time dobivamo sljedeće rezultate:

Riječ	Korijen	Riječ	Korijen
singt	sing	singen	sing
beliebt	belieb	beliebster	belieb
stören	stö	stöhnen	stöh
Kuß	Kuß *error!*	Küsse	Küss *error!*
Verlierer	Verli *error!*	Verlies	Verli *error!*
Maus	Mau *error!*	Mauer	Mau *error!*
stören	stö *error!*	Störsender	stö *error!*

Iz priloženog primjećujemo kako su nam samo prva tri retka točna. Primjećujemo kako riječ *Maus* i riječ Mauer imaju isti korijen iako to nije točno, jer ovdje govorimo o dvije različite

riječi s dva različita značenja. Isto vrijedi i za par Verlies i Verlierer. Kod riječi *Kuß* slovo ***ß*** ne smije biti korijen riječi, dok riječi *stören* i Störsender ne smiju imati znak ***ö*** kao korijen riječi. Slično je i s riječi *Küsse* gdje znak ***ü*** ne smije biti korijen riječi. Greške korjenovanja koje se pojavljuju u recima četiri i pet sprječavaju se algoritmom zamjene, dok su greške u recima šest i sedam više cijena koju plaćamo kako bi algoritam radio brže, jer smo prethodno definirali da naš algoritam provjerava osam različitih sufiksa umjesto svih 71 (iz navedenih sufiksa primjećujemo kako se oni ponekad međusobno uključuju, pri čemu zaključujemo da sufiksi vezani uz deklinaciju mogu biti sastavljeni od kombinacije sufiksa ***-e***, ***-s***, ***-n***, ***-t***, ***-em***, ***-er*** i ***-nd***) (Caumanns. 2008).

Cijeli prethodni proces korjenovanja smo napravili bez da smo uzeli u obzir sam kontekst ali i pravila koja nalaže njemačka lingvistika, koji nam zapravo daju dva razloga zbog čega se pojavljuju pogreške korjenovanja našeg korjenovatelja, a oni glase:

1. U njemačkom jeziku množina riječi se ponekad gradi tako da samoglasnik zamijenimo prijeglasom ili *Umlaut*-om (***ä***,***ö***,***ü***) i/ili zamjenjujemo znak ***ß*** sa ***ss***, što ne možemo postići isključivo uklanjanjem sufiksa. Isto pravilo vrijedi i za nepravilne glagole (*halten - hielt*).
2. Nizovi znakova koji skupa formiraju jedan zvuk (npr. ***-sch***, ***-ei***, i ***-ie***) budu našim algoritmom ponekad uklonjeni.

Kako bismo spriječili pogreške u korjenovanju koje smo imali, napravit ćemo zamjene prije samo uklanjanja nastavaka:

1. Svaki prijeglas je zamijenjen sa svojim pripadajućim samoglasnikom, te znak ***ß*** zamjenjujemo sa ***ss***.
2. Drugo slovo koje se pojavljuje dva puta za redom ćemo zamijeniti sa *.
3. ***-sch***, ***-ei***, i ***-ie*** ćemo zamijeniti s posebnim znakovima ***$***, ***§***, ***%***, ***&***

Prva zamjena ima svrhu točnog sastavljanja svih oblika množine. Prema izračunima autora Jörga Caumansa bilo kakvo pojavljivanje pogreški nakon ove zamjene je iznimno maleno (prema njegovoj tvrdnji 12 parova unutar liste od 50 000 imenica, primjeri za to su parovi *Stück* i *Stuck* ili *Eisbär* i *Eisbar* koji su imali jednak korijen nakon uklanjanja sufiksa). Druga i treća zamjena praktički služe očuvanju fiksnih nizova slova od njihove raščlambe. Svakako postojeći algoritam se može unaprijediti daljnjim zamjenama, ali je tu više riječ o prilagodbi korjenovatelja njemačkom jeziku (Caumanns, 2008).

2.2.2.2 Snowball- algoritam njemačkog korjenovatelja

U ovoj cjelini govorit ćemo isključivo o konkretnom algoritmu Snowbolla vezano uz korjenovanje njemačkog jezika. Ovaj algoritam razvio je Martin Porter sa svojim timom. Primjer njemačkog rječnika sa korjenovanim oblicima nalazi se u privitku pod poglavljem 5.3 Primjer rječnika sa korjenovanim oblicima pomoću Snowballa.

Na koji način on funkcionira? Prvenstveno se prepoznaju posebni znakovi (točnije ***ß***) i glasovi s prijeglasom (***ä***, ***ö***, ***ü***) te se s ***a***, ***e***, ***i o u*** i ***y*** definiraju kao samoglasnici. Prvo se zamijene ***ß*** sa ***ss*** te se postavljaju ***u*** i ***y*** između samoglasnika pisani velikim tiskanim slovima. R1 i R2 se definiraju kao prethodno navedenom poglavlju, te je nakon toga R1 način da niz prije njega sadrži barem tri slova. Definiraju se odgovarajuća slova koja smiju prethoditi nastavku ***-s***, a to su: ***b***, ***d***, ***f***, ***g***, ***h***, ***k***, ***l***, ***m***, ***n***, ***r*** *ili* ***t*** te se također definira lista slova koja smiju prethoditi nastavku *-st* kao ista lista slova, pri čemu se isključuje slovo r (*Snowball Tartarus*).

Prolazimo kroz 3 koraka Snowball algoritma:

1. korak: Traži se najduži među sljedećim sufiksima:
   a. ***em***, ***ern***, ***er***,
   b. ***e***, ***en***, ***es***,
   c. ***s***

   te ga obriši ako se nalazi u R1. Ako je odabran jedan od nastavaka ***e***, ***en***, ***e***, i kraju nastavka prethodi ***niss***, obriše se zadnje ***s***. Primjeri toga su:

   akcers ->acker

   armes -> arm

   bedürfnissen -> bedürfnis

2. korak: Traži se najduži sufiks među sljedećim grupama sufiksa:
   a. ***en***, ***er***, ***est***,
   b. ***st*** (mora prethoditi pravovaljano slovo iz liste slova, koje se smije pojavljivati prije nastavka ***st***, a koje smo definirali prethodno. Osim toga mora vrijediti za samo to slovo iz navedene liste, koje prije sebe mora imati barem tri slova)

te obriši ako se nalazi u R1. Primjeri bi bili:

derbsten -> derbst (1. korak) i *derbst -> derb* (2. korak) jer su zadovoljeni gore navedeni uvjeti drugog koraka.

3. korak: *d-sufiksi*

Ponovno se traži najduži sufiks među sljedećim skupinama te se primjenjuje određena radnja za svaku skupinu:

a. ***end***, ***ung***

obriši ako se nalazi u R2, a ako prethodi ***ig***, obriši ako je u R2 i ne prethodi mu slovo *e*.

b. ***ig***, ***ik***, ***isch*** **-**obriši ako se nalazi u R2 i ne prethodi mu slovo ***e***

c. ***lich***, ***heit*** **-***obriši* ako se nalazi u R2, te ako se prije njega pojavljuju ***en*** ili ***er***, obrisati ako se nalazi u R1

d. ***keit*** **-**obrisati ako se nalazi unutar R2, te, ako se prije njega pojavljuju ***lich*** ili ***ig***, obrisati ako je u R2.

Na posljetku ***U*** i ***Y*** ponovno vraćamo nazad u mala slova te se uklanjaju prijeglasi s ***ä***, ***ö***, ***ü***. Kod algoritma se nalazi u privitku (*Snowball Tartarus*).

2.2.2.3 UniNE korjenovatelj

Ovaj korjenovatelj je razvio Jacques Savoy 2006. godine na Sveučilištu u Neuchâtelu. Ovaj korjenovatelj ima dva načina korjenovanja:

1. agresivno korjenovanje i
2. jednostavno korjenovanje.

Jednostavnim pristupom korjenovanja nastoje se ukloniti nastavci koje tvore množinu riječi. Kada se izvrši zamjena preglasa, uklanja se jedan od nastavaka ***nen***, ***se*** i ***e*** prije jednog od nastavaka ***n***, ***r*** i ***s*** ili jednog od nastavaka ***n***, ***r*** i ***s*** koji se nalaze na kraju neke riječi. Bitno je ovdje napomenuti da se samo jedno od navedenih pravila može primijeniti u prvom koraku. (Weißweiler, Fraser, 2017)

S druge strane agresivno korjenovanje prolazi kroz mnogobrojne korake uklanjanja nastavaka, što uvijek ovisi o duljini riječi nad kojom se vrši proces korjenovanja. Za razliku od ostalih

korjenovatelja, UniNE koristi dvije grupe operacija za uklanjanje sufiksa, pri čemu se izvrši jedna od svake grupe. Također, uvjeti uklanjanja nastavaka ***s*** i ***st*** su zapravo vrlo slična onome algoritmu primijenjenom unutar Snowball korjenovatelja, pri čemu se definira popis konstantnih riječi koje su prihvatljive vezano uz uklanjanje nastavaka ***s*** i ***st*** te se trebaju pojaviti prije navedenih nastavaka kako bi se konstanta uklonila (Weißweiler, Fraser, 2017).

Ovdje je vrlo bitno napomenuti kako je korjenovatelju potrebno primijeniti jedanaest pravila ukoliko želi ukloniti nastavke za množinu te gramatičke nastavke (ovdje govorimo npr. o *-s* ili *-es* nastavcima kod genitiva neke imenice). Također množina njemačkog jezika ima dosta različitih nastavaka, a ne smijemo zaboraviti nit pojavljivanje prijeglasa ***ä***, ***ö***, ***ü*** u množini nekih riječi kao što bi npr bile *Apfel – Äpfel*, *Haus – Häuser* i druge (Savoy. 2006).

	Mean average precision					
IR Model \ Stemmer	Hungarian none	Hungarian light	Hungarian UniNE	German none	German UniNE	German Porter
doc=Okapi, query=npn	0.1957	**0.2988**	**0.3076**	**0.3552**	**0.3931**	**0.4058**
Prosit	0.1883	0.2905	0.2964	0.3464	0.3805	0.3934
doc=Lnu, query=ltc	0.1887	0.2913	0.2868	0.3357	0.3638	0.3793
doc=dtu, query=dtn	**0.1980**	0.2857	0.2900	0.3357	0.3671	0.3826
doc=atn, query=ntc	0.1794	0.2651	0.2755	0.3381	0.3653	0.3789
doc=ltn, query=ntc	0.1919	0.2556	0.2567	0.3184	0.3421	0.3573
doc=lnc, query=ltc	0.1616	0.2188	0.2153	0.2757	0.2983	0.3032
doc=ltc, query=ltc	0.1675	0.2207	0.2183	0.2575	0.2773	0.2891
doc=ntc, query=ntc	0.1713	0.2162	0.2079	0.2510	0.2649	0.2759
doc=bnn, query=bnn	0.1338	0.1748	0.1782	0.2430	0.2552	0.2637
doc=nnn, query=nnn	0.1326	0.1348	0.1256	0.1381	0.1419	0.1462

Slika 4: Rezultati točnosti korjenovanja koje je dobio Jacques Savoy u svojoj analizi korjenovatelja (Savoy. 2006)

Glavni problem ovog korjenovatelja možemo razmotriti na primjeru riječi *Adlers* (orlovi), pri čemu korjenovatelj kao korijen ove riječi pokazuje cijelu riječ, jer slovo *r* nije uključeno u popis valjanih slova koja se pojavljuju prije nastavka *s* kako bi se riječ mogla pravilno korjenovati (Weißweiler, Fraser. 2017).

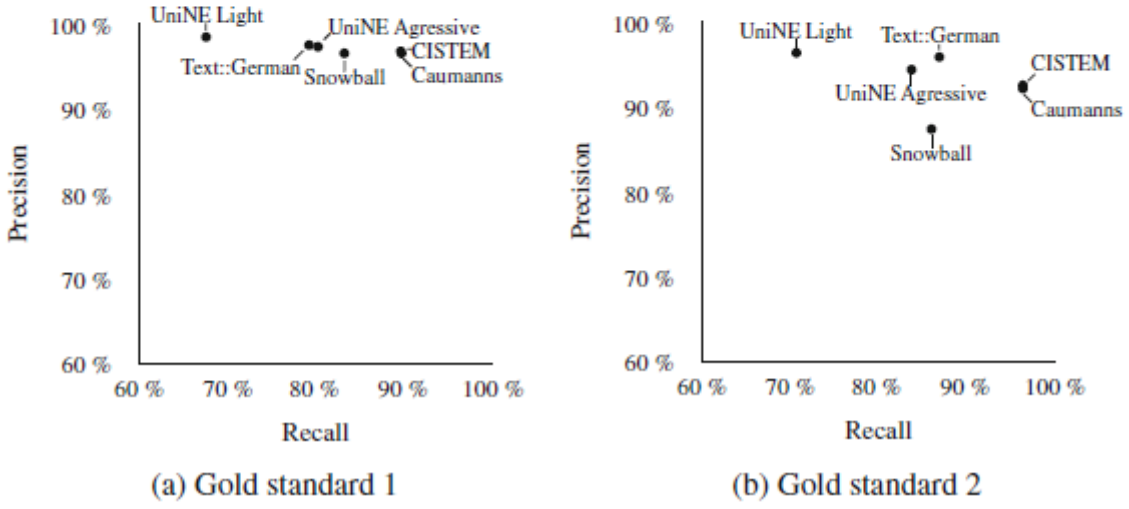
2.2.2.4 Text::Geramn korjenovatelj

Ovaj korjenovatelj jest zapravo modul Perl CPAN-a te ga je kao takvog razvio Ulrich Pfeifer 1996. godine na Tehničkom sveučilištu u Darmstadtu. Ono što razlikuje ovaj korjenovatelj od drugih jest uklanjanje prefiksa te koristi male liste prefiksa, sufiksa i korijena riječi kako bi prepoznao različite dijelove riječi. Autori tvrde da je unatoč njegovoj implementaciji u CPAN-u, ovaj korjenovatelj dao dobre rezultate tijekom testiranja (Weißweiler, Fraser. 2017).

Kao i prethodno navedeni korjenovatelj, i ovdje postoji problem s korjenovanjem riječi *Adlers*, jer sufiks ***ers*** nije uključen niti u jedan od njegovih popisa sufiksa. Autori su također uočili kako

ovaj korjenovatelj ne prebacuje velika pisana slova u mala, iako je na njihovom primjeru izvršio pravilno korjenovanje riječi, što je rezultiralo sa *adle* kao korijenom (Weißweiler, Fraser. 2017).

Gold standard 1						
Stemmer	Snowball	Text::German	Caumanns	UniNE Light	UniNE Aggressive	CISTEM
Precision	96.17%	97.56%	96.76%	**98.39%**	97.37%	96.83%
Recall	83.78%	79.29%	9.43%	67.69%	80.29%	**89.73%**
F1	89.55%	87.48%	92.95%	80.20%	88.01%	**93.15%**
Gold standard 2						
Stemmer	Snowball	Text::German	Caumanns	UniNE Light	UniNE Aggressive	CISTEM
Precision	85.89%	96.00%	92.26%	**96.43%**	94.50%	92.43%
Recall	86.61%	86.97%	96.17%	70.91%	83.81%	**96.45%**
F1	86.25%	91.27%	94.17%	81.72%	88.83%	**94.40%**

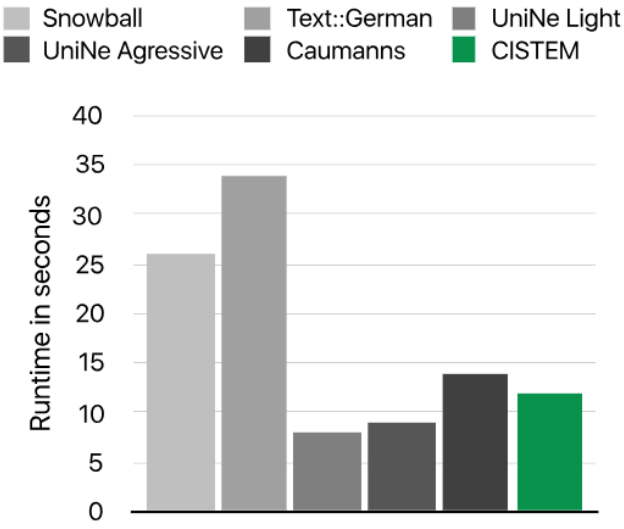


Slika 5:Usporedba točnosti svakog od krojenovatelja prema zlatnim standardima (Weissweiler, Fraser. 2017)

2.2.2.5 CISTEM korjenovatelj

Ovaj korjenovatelj je izgrađen kao poboljšanje Caumannsovog korjenovatelja, a primjenjuje uklanjanje zamjene slova ***z*** sa slovom ***x*** što je pridonijelo poboljšanju bez ikakvog daljnjeg utjecaja na druga pravila korjenovanja. Nadalje, za razliku od Caumannsovog korjenovatelja, uklanjanje predmetka ***ge*** iz riječi kao prefiksa, prije uklanjanja sufiksa riječi i pod uvjetom da je preostala riječ ima barem četiri slova, je dalo znatno bolje rezultate. Ono što su ovdje Weißweiler i Fraser su također primijetili jest da npr. zamjena glasa ***ei*** s posebnim znakom ***%*** čini riječ kraćom te dolazi do velikog odstupanja u točnosti algoritma. Osim toga također je uklonjena zamjena znakova ***ch*** sa ***§***, te je uvedena promjena vezana uz ograničenje dužine riječi tijekom uklanjanja nastavka ***nd*** koja je postavljena na barem šest preostalih slova nakon

uklanjanja nastavka umjesto prethodnih barem pet slova, što je dovelo ne samo do poboljšanja rada korjenovatelja nego i do pojednostavljenja algoritma, jer se sada nastavak ***nd*** uklanja u istom koraku kada i nastavci ***em*** i ***er***. Osim navedenog također je točno definiran slijed koraka, što nije slučaj kod Caumanns-ovog korjenovatelja (Weißweiler, Fraser, 2017).

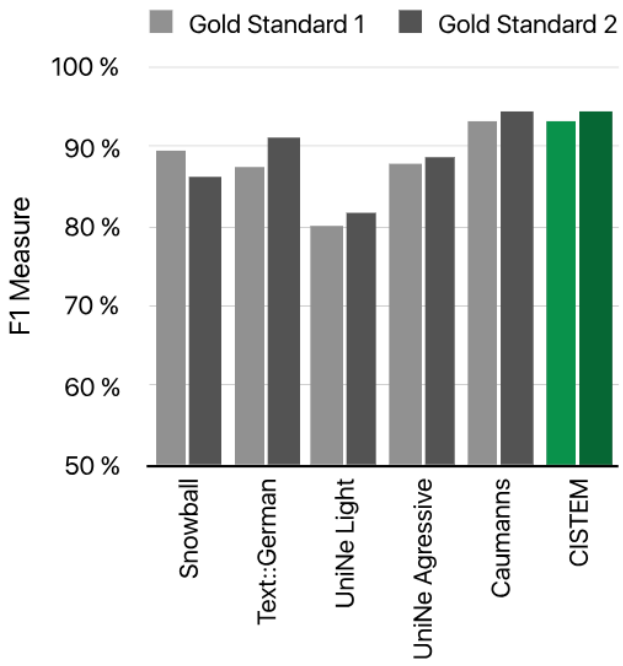


Slika 6: Usporedba brzine korjenovatelja koristeći jednu nit Perl 5.8.18 na Xeon x7560 2,26 GHz procesoru (Weißweiler, Fraser. 2017)

Koristeći set za evaluaciju 1[14] i set za evaluaciju 2[15] autori su izmjerili točnost svakog korjenovatelja te ih međusobno usporedili njihove rezultate.

14 Jedan dio dostupan u Privitku pod zasebnim poglavljem 5.7 Evaluacijski set 1 CISTEM korjenovatelja

15 Jedan dio dostupan u Privitku pod zasebnim poglavljem 5.8 Evaluacijski set 2 CISTEM korjenovatelja



Slika 7:Usporedba korjenovatelja po zlatnim standardima

2.2.3 Hibridni pristup korjenovanja

Hibridni korjenovatelji udružuju prethodna dva načina korjenovanja u jedan. Dok jedan računa statistiku mogućih točnih vrijednosti te ih vraća kao rezultat, drugi dio provjerava moguća rješenja prema pravilima morfologije jezika. Ovdje je riječ o pohrani iznimki u relativno male i jednostavne tablice koje ne zahtijevaju puno vremena za pripremu. Tijekom izvođenja se provjeri je li rezultat korjenovanja naveden unutar tabelice. Ako se takav rezultat ne pronalazi, vrši se uklanjanje afiksa prema morfološkim standardima ili pak lematizacija dane riječi (Feldweg. 1999). U ovom radu nisu detaljno analizirani hibridni korjenovatelji.

2.3 Obilježivači vrsta riječi

Obilježavanje vrsta riječi jest proces pridruživanja oznake vrste riječi svakoj sekvenci riječi ovisno o definiciji i kontekstu u kojem se ta riječ nalazi (odnos sa susjednim i srodnim riječima unutar fraze, rečenice ili paragrafa). Prethodno je potrebno odvojiti sve dijakritičke znakove od riječi uzimajući u obzir postojanje dijakritičkih znakova kao dijelova riječi (na primjer oznaka itd., npr. ili pak t. d.) (Jurish. 2003).

Postoji više načina obilježavanja vrsta riječi:

1. Obilježivači vrsta riječi bazirani na pravilima, koja su zadana ručno kako bi se raspoznala dvosmislenost oznaka

2. Obilježivači vrsta riječi bazirani na HMM-u ili VMM-u odabira niza oznaka koje maksimiziraju produkt vjerojatnosti riječi i vjerojatnost sekvenci oznake. Ovakvi obilježivači su učeni na ručno označenim korpusima u kombinaciji s različitim stupnjevima n-grama koristeći interpolacije brisanja i sofisticirane modele za nepoznate riječi.
3. Ostali pristupi obilježavanja vrsta riječi obuhvaćaju metode dinamičkog programiranja (DeRoseova metoda i Churchova metoda), Viterbi algoritam, Baum-Welchov algoritam, hibridni pristupi obilježavanju vrsta riječi (ručno pisana pravila u kombinaciji s vjerojatnosnim modelima) itd.

2.3.1 Obilježivači vrsta riječi bazirani na HMM-u ili VMM-u

Kao što smo prethodno napomenuli ovakvi modeli zahtijevaju učenje na ručno označenim korpusima. Njihova prednost leži u njihovoj mogućnosti rješavanja mnogih problema dvosmislenosti, jer skoro bilo koji problem procesiranja govora ili jezika se može predstaviti kao „danih *N* izbora za neki dvosmisleni unos, odaberi onaj koji ima najveću vjerojatnost“ (Jurafsky, Martin. 2007).

2.3.1.1 Obilježivači vrsta riječi bazirani na HMM-u

Korištenjem ovog modela govorimo o posebnom slučaju Bayesove klasifikacije koja glasi:

$$\hat{\boldsymbol{t}}_{\boldsymbol{1}}^{\boldsymbol{n}} = \boldsymbol{argmax}_{\boldsymbol{t}_{\boldsymbol{1}}^{\boldsymbol{n}}} \boldsymbol{P}(\boldsymbol{t}_{\boldsymbol{1}}^{\boldsymbol{n}} | \boldsymbol{w}_{\boldsymbol{1}}^{\boldsymbol{n}})$$

To znači da gledamo sve moguće nizove oznaka. Iz tog seta želimo odabrati onu oznaku koja ima najveću vjerojatnost s obzirom na niz *n* riječi $\boldsymbol{w}_{\boldsymbol{1}}^{\boldsymbol{n}}$. Kod drugih riječi želimo od svih nizova od *n* oznaka $\boldsymbol{t}_{\boldsymbol{1}}^{\boldsymbol{n}}$ jedan niz oznaka takvih da je vjerojatnost $\boldsymbol{P}(\boldsymbol{t}_{\boldsymbol{1}}^{\boldsymbol{n}} | \boldsymbol{w}_{\boldsymbol{1}}^{\boldsymbol{n}})$ najveća. Koristimo notaciju ^ kako bismo označili „našu pretpostavku točnih nizova oznaka“. Funkcija $argmax_x f(x)$ znači „takav *x* u kojem funkcija *f(x)* postiže svoj maksimum“. Cjelokupna jednadžba znači da od svih sekvenci oznaka duljine *n*, želimo dobiti točan niz znakova $\boldsymbol{t}_{\boldsymbol{1}}^{\boldsymbol{n}}$ koji postiže maksimum na svojoj desnoj strani. Međutim i dalje nam preostaje problem kako za dani niz oznaka $\boldsymbol{t}_{\boldsymbol{1}}^{\boldsymbol{n}}$ i niz riječi $\boldsymbol{w}_{\boldsymbol{1}}^{\boldsymbol{n}}$ direktno izračunati $\boldsymbol{P}(\boldsymbol{t}_{\boldsymbol{1}}^{\boldsymbol{n}} | \boldsymbol{w}_{\boldsymbol{1}}^{\boldsymbol{n}})$ (Jurafsky, Martin. 2000).

Kako bismo uspješno primijenili Bayesovu klasifikaciju, potrebno je i jednostavnije primijeniti Bayesovo pravilo koje rastavlja uvjetnu vjerojatnost *P(x|y)* na tri druge vjerojatnosti:

$$\boldsymbol{P(x|y) = \frac{P(y|x)P(x)}{P(y)}}.$$

Uvrštavanjem druge jednadžbe u prvu (fusnota 18) dobivamo sljedeći iskaz:

$$\boldsymbol{\hat{t}_1^n = argmax_{t_1^n} \frac{P(w_1^n \vee t_1^n)P(t_1^n)}{P(w_1^n)}}$$

Ovu jednadžbu (ovdje stavi fusnotu 18 možemo pojednostaviti izbacivanjem nazivnika $\boldsymbol{P(w_1^n)}$, jer je konstantan za niz $\boldsymbol{w_1^n}$ pa pojednostavimo u:

$$\boldsymbol{\hat{t}_1^n = argmax_{t_1^n} P(w_1^n \vee t_1^n)P(t_1^n)}$$

Ova jednadžba nam govori kako se niz oznaka s najvećom vjerojatnosti $\boldsymbol{\hat{t}_1^n}$ za dani niz znakova $\boldsymbol{w_1^n}$ se može izračunati uzimanjem produkta vjerojatnosti za svaku sekvencu te odabiranjem niza oznaka za koji je taj produkt najveći. Ovdje govorimo o vjerojatnosti prethodnog niza znakova $\boldsymbol{P(t_1^n)}$ i vjerojatnosti niza znakova riječi (Jurafsky, Martin. 2000).

Međutim, unatoč ovom pojednostavljenju i dalje je teško direktno izračunati vjerojatnost. Stoga uvodimo pretpostavke koje potpomažu pojednostavljenju. Prva pretpostavka nam govori kako vjerojatnost pojavljivanja riječi ovisi isključivo o vlastitom obilježivaču vrsta riječi, tj. da je neovisna o drugim riječima oko nje te oznaka oko te riječi:

$$\boldsymbol{P(w_1^n|t_1^n) \approx \prod_{i=1}^{n} P(w_i \vee t_i)} \quad (16)$$

Druga pretpostavka nam govori kako vjerojatnost pojavljivanja oznake ovisi isključivo o prethodnoj oznaci (fusnota 19):

$$\boldsymbol{P(t_1^n) \approx \prod_{i=1}^{n} P(t_i \vee t_{i-1})}.$$

Uvrštavanjem ove dvije pretpostavke u prethodno pojednostavljenu jednadžbu dobivamo sljedeću formulu po kojoj obilježivač bigrama pretpostavlja sekvencu oznaka koje imaju najveću vjerojatnost (Jurafsky, Martin. 2000):

16 Jurafsky, D., Martin, J. H. (2000). *Speech and language Processing: An introduction to natural Language processing, Computational Linguistics and Speech Recognition*. Prentice Hall

$$\hat{t}_1^n = argmax_{t_1^n} P(t_1^n \vee w_1^n) \approx argmax_{t_1^n} \prod_{i=1}^{n} P(w_i \vee t_i) P(t_i \vee t_{i-1}).$$

Ova jednadžba sadrži dvije vrste vjerojatnosti, vjerojatnost tranzicije oznaka i vjerojatnost pojavljivanja neke riječi. Vjerojatnost prijelaza oznake $P(t_i \vee t_{i-1})$ predstavlja vjerojatnost pojavljivanja prethodne oznake t_{i-1}. Ovu formulu najbolje možemo provjeriti na temelju određenih i neodređenih članova uz pridjev u njemačkom jeziku. Svakako postoje situacije kada se ne niti jedan od navedenih ne pojavljuje, već imamo situacije gdje se jednostavno imenica i njen pridjev pojavljuju isključivo u takvom obliku. Nažalost gore navedena formula nam neće vratiti previše precizne rezultate, pa je tako ovdje potrebno odrediti izračun omjera brojača koji će nam izračunati vjerojatnost pojavljivanja (određenog ili neodređenog) člana u kombinaciji s pridjevom, što znači da će naš izračun izgledati ovako:

$$P(t_i|t_{i-1}) = \frac{C(t_i-1,t_i)}{C(t_i-1)}.$$

Kada govorimo o vjerojatnosti istoznačnih riječi $P(w_i \vee t_i)$, uz pretpostavku da vidimo oznaku, predstavlja vjerojatnost točnog povezivanja oznake s riječi. Uzmemo li npr. oznaku VAINF iz seta oznaka TIGER korpusa koja predstavlja infinitivni oblik pomoćnih glagola (u njemačkom jeziku najčešći su pomoćni glagoli *sein* -biti i *werden*-postati koji se pojavljuju u tom obliku). Za takav zadatak koristili bismo sljedeći izračun:

$$P(w_i|t_i) = \frac{C(t_i \vee w_i)}{C(t_i)},$$

te dobili konkretan izračun pojavljivanja svakoga od njih koji bi bio poprilično velik. Tako smo definirali obilježavanje pomoću skrivenih Markovljevih modela i Bayesovog teorema (Jurafsky, Martin. 2000).

2.3.1.2 Obilježivači vrsta riječi bazirani na VMM-u

Markovljev model varijabilne memorije, (eng. *Variable Memory Markov Model*) definiramo kao aproksimaciju neograničenog Markovljevog reda izvora. Takav model može sistematično inkorporirati statičke (nultog reda) i dinamičke (višeg reda) informacije, u međuvremenu posjedujući mogućnost izmjene modela pod utjecajem budućih promatranja. Ovaj pristup obilježavanja riječi je jednostavan za implementaciju, a algoritmi učenja su računski učinkoviti (Schütze, Singer. 1994).

Algoritam učenja je baziran na minimizaciji statističkog predviđanja pogreške Markovljevog modela izmjerenog pomoću trenutnog KL odstupanja[17]. Memorija se proširuje u trenutku kada je promjena dovoljno bitna kako bi cijelo statističko predviđanje stohastičkog modela bilo dovoljno dobro. Kako bi ovaj algoritam funkcionirao definiramo $\sum$ kao konačan abeceda te označimo $\sum^*$ kao set svih nizova znakova nad $\sum$. Niz znakova *s* nad $\sum^*$ duljine *n* definiramo kao $s = s_1 s_2 \ldots s_n$. Duljinu niza znakova definiramo sa |*s*|, a veličina abecede $\sum$ kao $|\sum|$. $\boldsymbol{Prefix(s) = s = s_1 s_2 \ldots . s_{n-1}}$ predstavlja najduži prefiks niza znakova *s*, dok *Prefix*(s)* predstavlja set svih prefiksa od *s* uključujući prazan niz znakova. Također $\boldsymbol{Suffix(s) = s_2 \ldots s_n}$, dok nam *Suffix*(s)* predstavlja set svih sufiksa od *s*, a *e* je prazan niz znakova. Vjerojatnost *P* nad $\sum^*$ je pravovaljana ako je *P(e)* = 1, a za svaki niz znakova *s* $\sum_{\sigma \epsilon \sum} \boldsymbol{P(s\sigma) = P(s)}$. Za svaki set riječi *S* koji ne sadrži prefikse $\sum_{s \epsilon S} \boldsymbol{P(s) \leq 1}$, naročito za svaki cijeli broj $n \geq 0$, $\sum_{s \epsilon \sum^n} \boldsymbol{P(s) = 1}$ (Schütze, Singer. 1994).

Stablo predviđanja sufiksa *T* nad $\sum$ je stupnja$|\sum|$, čiji su rubovi označeni simbolima iz $\sum$ (naše abecede) tako da kod svakog unutarnjeg čvora postoji najviše jedan izlazeći rub označen svakim simbolom. Čvorovi stabla su označeni parovima (s , γ_s), pri čemu *s* predstavlja niz znakova kao put kroz stablo od tog čvora pa sve do samog korijena stabla, a $\gamma_s : \sum \rightarrow [0,1]$ predstavlja funkciju vjerojatnosti izlaza od *s*, koji zadovoljava $\sum_{\sigma \epsilon \Sigma} \boldsymbol{\gamma_s(\sigma) = 1}$. Osim navedenog, potrebno je također uzeti u obzir i vjerojatnosni konačni automat, koji ima svoje zasebne formule za računanje vjerojatnosti i djelovanja na abecedu $\sum$, te na Markovljeve procese reda L (eng. *Markov processes of order L*)[18]. To znači da broj stanja *n* je znatno manji od $|\sum|^L$, što znači da rijetko koje stanje ovog automata ima duže pamćenje, dok većina stanja ima kraće pamćenje. Sve navedeno povezujemo pod nazivom Markovljev proces varijabilne memorije ili skraćeno VMM (eng. *variable memory Markov process*). U slučaju Markovljevog procesa stupnja *L* identitet stanja je poznat te učenje takvih procesa svodi na aproksimiranje funkcije vjerojatnosti izlaza (Schütze, Singer. 1994).

Sada kada smo ukratko prošli način djelovanja VMM-a, postavljamo pitanje kako je koristiti za obilježavanje riječi. Sada kada smo naučili VMM na označenom korpusu, vjerojatnost svake specifične riječi koja pripada određenoj klasi oznaka je pretpostavljena koristeći maksimalnu

17 Kullback-Leiblerovo odstupanje je mjera kako se jedna distribucija vjerojatnosti razlikuje od druge, očekivane raspodjele vjerojatnosti.

18 Ovdje govorimo o stohastičkom modelu duljine *L* koji opisuje niz mogućih događaja u kojima vjerojatnost svakog događaja ovisi samo o stanju koje je postignuto u prethodnom stanju.

procjenu vjerojatnosti svakog individualnog brojača riječi. Vjerojatnosti stanja i prijelaza Markovljevog modela su određeni algoritmom učenja te sue vjerojatnosti izlaza oznaka (eng. *tag output probabilities*) pretpostavljene iz brojača riječi (statičke informacije se nalaze zapisane u korpusu koji smo koristili za učenje). Uvodimo formule koje smo definirali u prethodnom podnaslovu pod fusnotama 18 i 19 osim zadnje koja nam kasnije služi za utvrđivanje statičkih parametara (Schütze, Singer. 1994).

Pretpostavimo li da je memorija Markovljevog modela M, tada je $P(t_i \mid t_{1,i-1})$ izračunat na temelju $P(t_i \mid S_{i-1}, M)$, pri čemu je $\boldsymbol{S_i = r(\varepsilon, t_{1,i})}$ budući da je dinamičnost sekvenci predstavljena vjerojatnostima prijelazima korespondirajućeg automata. To znači da su stoga oznake $t_{1,n}$ za sekvence riječi $w_{1,n}$ odabrani pomoću sljedeće jednadžbe Viterbi algoritma:

$$\boldsymbol{T_M(w_{1,n}) = argmax_{t_{1,n}} \prod_{i=1}^{n} P(t_i \vee S_{i-1}, M) P(w_i \vee t_i)},$$

$\boldsymbol{P(w_i \vee t_i)}$ indirektno pretpostavljamo iz $P(t_i \mid w_i)$ koristeći Bayesov teorem:

$$\boldsymbol{P(w_i|t_i) = \frac{P(t_i|w_i)P(w_i)}{P(t_i)}}.$$

$\boldsymbol{P(w_i)}$ su konstantne za danu sekvencu $\boldsymbol{w_i}$ te se stoga mogu zanemariti tijekom traženja maksimuma funkcije. Nakon toga izvršavamo maksimalnu procjenu vjerojatnosti $\boldsymbol{P(t_i)}$ računajući relativnu frekvenciju od t_i.

Kao što sam napomenuo, za kraj nam ostaje procjena statičkih parametara, koju ćemo izračunati pomoću već navedene formule:

$$\boldsymbol{P(t^i|w^j) = \frac{C(t^i, w^j)}{C(w^j)}},$$

gdje $C(t^i, w^j)$ predstavlja broj koliko puta je t^i je označen sa w^j u tekstu koji koristimo za učenje, a $C(w^j)$ predstavlja broj koliko puta se w^j pojavljuje u treniranom tekstu. Iako dobra ideja, ipak je potrebna glatkoća funkcije, jer će bilo koji drugi tekst sadržavati druge riječi pa će nam stoga broj $C(w^j)$ biti jednak nuli. Jedno od rješenja ovom problemu jest tzv. *Good-Turing*-ovog procesa:

$$\boldsymbol{P(t^i|w^j) = \frac{C(t^i, w^j) + 1}{C(w^j) + I}},$$

gdje $\boldsymbol{I}$ predstavlja broj oznaka. Ponekad u nekim situacijama ovaj pristup neće biti najbolje rješenje te će nakon modifikacije (ovdje uzimamo u obzir vjerojatnost konverzije oznaka, za koju ćemo raditi zasebna predviđanja i izračune) njen konačan izgled biti sljedeći:

$$\boldsymbol{P(t^i|w^j) = \frac{C(t^i, w^j) + \sum_{k_1 \epsilon T_j} P_{1m}(k_1 \rightarrow i)}{C(w^j) + \sum_{k_1 \epsilon T_j, k_2 \epsilon T} P_{1m}(k_1 \rightarrow k_2)}}$$

gdje će naš $\boldsymbol{T_j}$ predstavlja set oznaka koje w^j sadrži u svom setu za učenje, a $\boldsymbol{T}$ je naš početni set svih oznaka. Ovaj pristup ima mali učinak na predviđanja bazirana na velikim brojačima koje smo koristili za utvrđivanja, poštuje se razlika između zatvorenih i otvorenih klasa riječi (vjerojatnost konverzije zatvorene klase je jednaka nuli te, samim time, glatkoća funkcije ne utječe na nju) i implementirali smo početno poznavanje vjerojatnosti konverzije različitih klasa oznaka (Schütze, Singer. 1994).

2.3.2 Obilježivači vrsta riječi bazirani na ručno pisanim pravilima/gramatikama

Ova vrsta obilježivača riječi ne koristi statistički pristup, već se bazira na ručno pisanim pravilima ili gramatikama. Jedan od proširenijih pristupa obilježavanju vrsta riječi je Brillov pristup označavanju vrsta riječi čiji nadzirani model učenja radi na principu učenja podataka na točno označenom tekstu, nakon čega se novi tekst bez prethodne anotacije provlači kroz početni program obilježavanja, što nam daje točno označavanje. Oznake se dodjeljuju po principu najvjerojatnije oznake koja se nalazi na našem leksičkom popisu, što je odlučeno na korpusu koji smo učili. Prednost ovog pristupa u odnosu na prethodno opisane pristupe jest kompaktnost, jer se sastoje od par stotina pravila koja se mogu jednostavno pregledati (kod prethodnih modela govorimo o tisućama kontekstualnih vjerojatnosti) te otpornost na učinke pretjerane prilagodbe modela na podatke za učenje. Pošto ovaj pristup nije zastupljen kod obilježavanja vrsta riječi njemačkog jezika, neću opisivati algoritam, ali je Brillov pristup polaznica ostalim algoritmima koje ću opisati u nastavku.

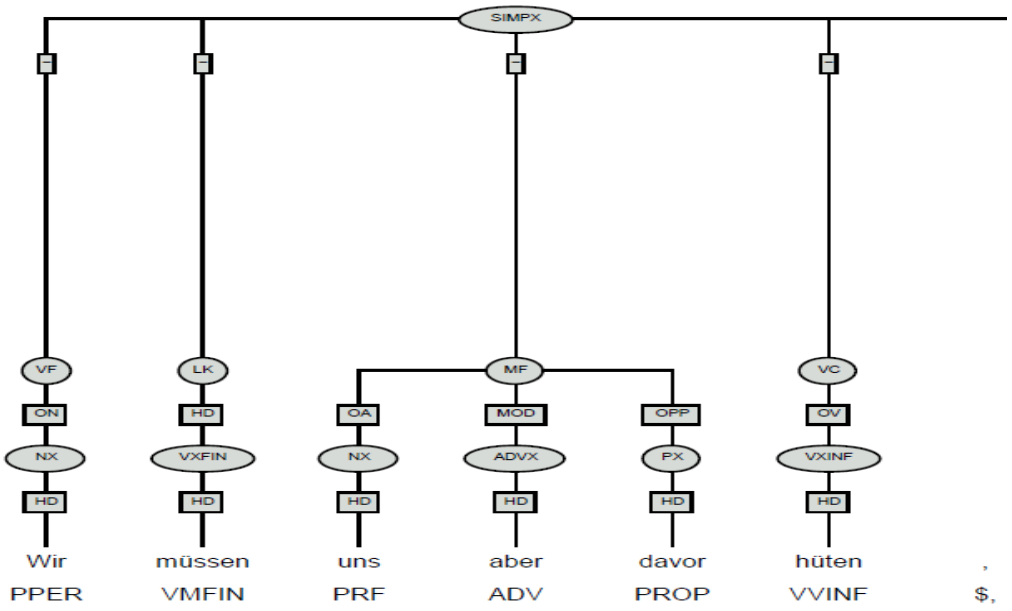
Vezano uz ovu vrstu obilježivača riječi najviše se ističu obilježivači vrsta riječi integrirani unutar TüBA-D/Z i TIGER korpusa, Stanfordov obilježivač vrsta riječi (model napravljen za njemački jezik).

2.3.2.1 Obilježivač riječi TüBA-D/Z korpusa

Algoritam obilježavanja riječi ovog alata je baziran na onomu iz Verbmobil Treebank-a koji razlikuje 4 razine sintaktičke konzistencije: leksička razina, razina topografskih područja, frazalna razina i razina klauzule. Prema principu primarnog poretka klauzule je popis

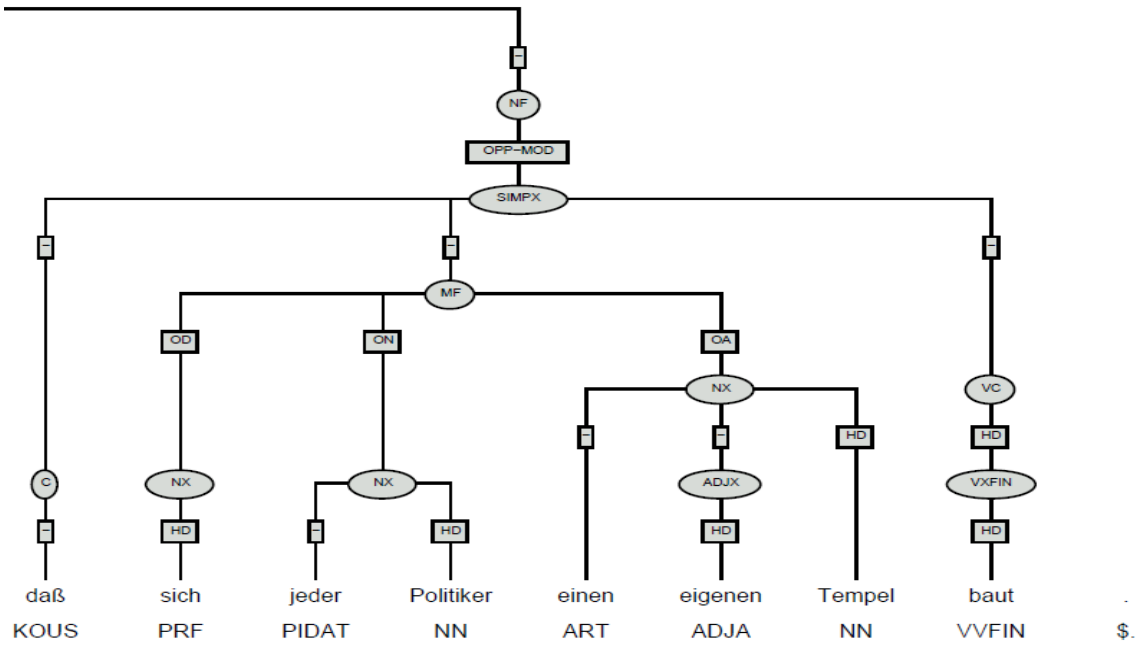
topoloških polja, koje karakterizira regularnost redoslijeda riječi među različitim tipovima klauzula njemačkog jezika, koje su široko prihvaćene među opisnim lingvistima njemačkog jezika. Uz konstantu strukturu, označena stabla sadrže rubne oznake između čvorova. Te rubne oznake između čvorova šifriraju gramatičke funkcije kao relacije između fraza te šifrira razliku između početka rečenice (koju ovdje definiramo kao glavu HD) i ostalih dijelova koji nisu početak rečenice (-) kao odnose unutar neke fraze (Telljohann, Hinrichs, Kübler. 2004).

Uzmemo li na primjer sljedeću rečenicu: *Wir müssen uns aber davor hüten, dass sich jeder Politiker einen eigenen Tempel baut.* (na hrv. „Moramo prijevremeno spriječiti kako svaki političar ne bi sebi sagradio vlastiti hram. “)



Slika 8: Lijeva grana TüBa-D/Z stabla

Rečenica, koju ćemo prema TüBA-D/Z označiti sa SIMPX, je grupirana na sljedeća topološka područja: početno područje označimo s VF, lijevo od toga imamo lijevi držač rečenice, kojeg označavamo s LK, srednje područje rečenice MF, drugi glagol kojeg označavamo s VC te završno područje NF (stajemo na zarezu, jer je drugi dio praktički umetnuta rečenica). Svršeni glagol nam pokazuje “glavu” klauzule (Telljohann, Hinrichs, Kübler. 2004).



Slika 9: Desni grana TüBa-D/Z stabla

Gramatički odnosi označeni u stablu su: subjekt (ON), objekt u akuzativu (OA), objekt u dativu (OD), glagolski objekt (OV), prijedložni objekt (OPP), modifikator prijedložnog objekta (OPP-MOD) te modifikator (MOD) (Telljohann, Hinrichs, Kübler. 2004).

Sintaktička i semantička dvosmislenost tretira se u smislu podređenosti, koja se oslanja na načelo visoke vezanosti i koristi nedefiniranu oznaku čvora. Kada je dvosmislenost moguća, odabire se nedvosmislena oznaka, kao što smo to u prethodnom primjeru uzeli oznaku OPP-MOD. Ako se dvosmislenost nikako ne može razriješiti, modifikator dvosmislenosti dobiva oznaku MOD te se priključuje na najviši mogući čvor (Telljohann, Hinrichs, Kübler. 2004).

2.3.2.2 Obilježivač riječi TIGER korpusa

TIGER korpus obuhvaća više od 40 000 sintaktički označenih rečenica njemačkog jezika baziranih na različitim novinskim člancima njemačkih vjesnika. Ovdje govorimo o oko 800 000 riječi koji se svakodnevno nadopunjuje te, prema novijim mjerenjima, sadrži čak 1 500 000 riječi u približno 80 000 riječi (Telljohann, Hinrichs, Kübler. 2004).

Kada govorimo o obilježavanju vrsta riječi, TIGER korpus primjenjuje 25 oznaka čvorova u 50 gramatičkih rubnih oznaka za sintaktičke kategorije za razliku od prethodnog obilježivača korpusa (TüBA-D/Z) koji obuhvaća 25 oznaka čvorova te 36 rubnih oznaka za sintaktičke kategorije. Pored standardnih oznaka gramatičkih funkcionalnosti unutar rečenice (kao što su recimo SB – *Subjekt* ili OA- *Akkusativ Objekt*), ovdje govorimo o proširenju pojmova

gramatičkih funkcija unutar rečenice kao što bi na primjer to bili RE (elementi koji se često ponavljaju unutar rečenice) te RS (upravni govor) (Smith. 2003).

Promotrimo rezultate dobivene na sljedećem konkretno primjeru rečenice: *Bei den Gesprächen in London würden zudem Hilfen für Osteuropa und die frühere Sowjetunion behandelt.* (Tijekom rasprave u Londonu, dotaknuta je tema potpore zemljama istočne Europe i Sovjetskom savezu.) Rezultate analize ove rečenice predočeni su unutar privitka (lijeva i desna strana stabla zasebno radi preglednosti) pod pripadnim naslovom[19].

Set oznaka TIGER korpusa je baziran na Stuttgart-Tübinger Tagset (STTS) s manjim promjenama. STTS dijeli vrste riječi njemačkog jezika na 11 glavnih kategorija nad kojima se kasnije vrši dodatna potpodjela. Na primjeru, te podklase bi izgledale ovako:

Part of Speech	Verb Type	Finiteness
V *Verb*	A *Auxiliar*	FIN *finit*
		INF *infinit*
		IMP *imperativ*
		PP *Partizip Perfekt*
	M *Modal*	FIN *finit*
		INF *infinit*
		PP *Partizip Perfekt*
	V *Voll*	FIN *finit*
		INF *infinit*
		IZU *Infinitiv mit zu*
		IMP *imperativ*
		PP *Partizip Perfekt*

Slika 10: Podklase glagola u STTS-u

To bi značilo da, npr., oznaka VMFIN predstavlja modalni glagol u njegovom svršenom obliku, tj. neinfonitvnom, konjugiranom obliku. Potraga pripadnosti vrsti riječi određenoj kategoriji unutar TIGER korpusa korištenjem izraza u TIGER korpusu poznato je pod nazivom *node description* (opis čvora). Najjednostavniji opis čvora se sastoji od izraza znanog kao ograničenje značajki (eng. *feature constraint*). Najjednostavnije ograničenje značajki sastoji se od jednog značajka-vrijednost para, pri čemu su značajka i vrijednost odvojeni znakom

19 5.4 Primjer stabla TIGER korpusa

jednakosti (npr. [pos = “ ART“] ili [pos = “NN“], što znači da opis čvora nam govori kako čvor ima vrijednost ART- *Artikel* za značajku pos-part-of-speech) (Smith, 2003).

Sastavne kategorije su kodirane u čvorovima oznaka, pri čemu su frazalne kategorije predstavljene nezavršnim čvorovima, dok završni čvorovi predstavljaju vrste riječi. Stabla sastavnih struktura TIGER korpusa su riješena svih redundantnih grana. Razlika između argumenata i dodataka nije klasificirana unutar strukture sastavnice, već je izražena pomoću sintaktičkih funkcija. Samom eliminacijom redundantnih struktura, koje možemo vrlo lako pronaći te nisu zapravo potrebne za procesiranje zahtjeva, rezultira stablima koje je mnogo lakše iščitati na ekranu. Konkretni primjeri eliminacije redundancije bi bili zapravo samo uklanjanje i negrananje NP čvora te kodiranje prijedložnih izraza bez odvojenog NP čvora za ono što u njemačkom jeziku zovemo *Nominalphrase* koju dominira prijedložni izraz, što čini taj set predvidljivim.

TIGER treebank također poznaje sekundarne rubove (eng. *secondary edges*) koje pozicioniramo između dva čvora koja nisu u direktnom odnosu, te označavaju sintaktički odnos ta dva čvora. Na konkretnom primjeru rečenice: „*Sie entwickelt und druckt Verpackungen und Etiketten.*“ (Ona razvija i stvara pakiranja i printa etikete.) početna imenica *Sie* funkcionira ne samo kao subjekt prvog dijela rečenice, već i kao subjekt drugog dijela, te tu uvodimo sekundarni rub. Tu smo se već dotakli područja sintaktičkih odnosa (točnije sekundarnih sintaktičkih odnosa na temelju prethodnog primjera), koje su definiran na rubu nastalog vezom čvora-roditelja i čvora-djeteta. Na konkretnom primjeru: „*Der Parteitag der SPD begann am Mittwoch um 15:15 Uhr.*“(Dan stranke SPD započeo je srijedom u petnaest sati i petnaest minuta.) gdje imamo NP (*Nominalphrase*) ili imenični izraz koji unutar sebe sadrži drugi imenični izraz, koji kasnije ima ulogu atributa u genitivu (u oznaci AG), sintaktička struktura je definirana rubnom oznakom veze čvor-roditelj i čvor-dijete (Smith, 2003).

2.3.3 Hibridni obilježivači vrsta riječi

U prethodnim poglavljima sam govorio o principu rada obilježivača vrsta riječi baziranog na HMM ili VMM modelu. Iako su se ti modeli pokazali veoma uspješnima, nažalost i dalje dolazi do visokog postotka pogreške, što varira od 14% na više. Iako su se statistički modeli pokazali kao bolji izbor u odnosu na one bazirane na ručno zadanim pravilima parsiranja, još uvijek imamo problem što dijelovi baznog korpusa tijekom učenja mogu biti optimizirani ili se učenjem gubi dio sintaktičkih veza unutar rečenice postavljanjem vjerojatnosti na sve mogućnosti uloge riječi unutar rečenice, tim više za jezike poput njemačkog koji sam po sebi

daje dodatne mogućnosti poput pretvaranja participa perfekta glagola u pridjev koji provlačimo kroz sva četiri padeža njemačkog jezika za svako lice svakog roda i broja. Kako ne bi dolazilo do sličnih problema te kako bi se poboljšali rezultati stvoren je ovaj koncept, koji kombinira metode učenja iz teksta s procjenom parametara iz teksta koji je prethodno označen obilježivačima vrsta riječi i njihovim ulogama unutar rečenice.

U ovoj situaciji početni prijelazi i utjecaji simbola zamijenjeni sa slijedom oznaka i pristranost oznaka u određenim vremenskim intervalima iz relativno malog, prethodno označenog, teksta. Ti vremenski intervali ili frekvencije se uzimaju kao aproksimacija vjerojatnosnog modela te se postiže bolja glatkoća funkcije u manjem broju iteracija. Ovim pristupom početni postotak pogreške od najmanje 14% se smanjuje čak na najmanje 3% pa sve do 8% ili 9% (ovisno o autoru, algoritmu učenja i baznom tekstu na kojem je izvršeno testiranje, što prema Feldwegovim testiranjima na Xerox obilježivaču razvijenog na Sveučilištu u Stuttgartu iznosi u rasponu od 3.16% do 7,29% 1999. godine) (Senrich, Kuntz. 2014).

Od novijih implementacija ovog pristupa obilježavanja vrsta riječi obrađen je na Sveučilištu u Zürichu (Pro3GresDE) te na Berlin-Brandenburgische Akademie der Wissenschaften (matematički algoritam programa za koji bi mi bila potrebna licenca, nismo analizirali) (Senrich, Kuntz. 2014).

2.3.3.1 Hybridni Pro3GresDE parser

Ovdje govorimo o robusnom i brzom bi-leksičkom parseru prilagođenom pravilima njemačkog jezika, koji koriti hibridnu arhitekturu ručno pisanih pravila funkcionalne ovisne gramatike sa statističkom leksičkom razdiobom dobivenom iz Penn Treebank-a. Metoda razjašnjavanja proširuje pristup PP-privrženosti na sve glavne vrste ovisnosti. Formula vjerojatnosti prilagodbe za sintaktički odnos *R* na udaljenost *dist*, s obzirom na to da se leksički predmeti *a* i *b* izračunavaju pomoću MLE procjene, uključujući nekoliko povratnih razina, bi izgledala ovako:

$$\boldsymbol{P(R, dist|a, b) \cong p(R|a, b) \cdot p(dist|R) = \frac{f(R,a,b)}{\sum_{i=1}^{n} f(R_i,a,b)} \cdot \frac{f(R,dist)}{fR}.}$$

To omogućuje parseru rezanje tijekom parsiranja te kao rezultat vraća moguće analize po njihovoj vjerojatnosti koje su prihvatljive po pravilima gramatike (Senrich, Kuntz. 2014).
Kada na to dodamo prilagodbu parsera na relativno slobodan redoslijed riječi u rečenici njemačkog jezika, uključuju se morfološka i topološka pravila gramatike kako bi se bolje identificirale granice imenskog predikata (vrši se pomoću GERTWOL sustava morfološke

analize koji je inače uključen unutar TüBa-D/Z korpusa), te dolazi do znatno boljih rezultata, zbog smanjenja redundancije i istoznačnosti kroz djelovanje morfološkog analizatora. Istoznačnost se drastično smanjuje unutar imenskog predikata te između subjekta i glagola upotrebom unaprijed definiranih gramatičkih pravila njemačkog jezika (Senrich, Kuntz. 2014).

Unatoč uspješnoj adaptaciji, algoritam i dalje ima problema s objektima u genitivu, imenskim predikatom u vokativu ili priložnim funkcijama.

3 Praktična usporedba alata za procesiranje njemačkog jezika

3.1 Usporedba obilježivača vrsta riječi

U ovom dijelu uspoređujemo točnost dvaju obilježivača vrsta riječi, i to Pro3GresDE te onog integriranog unutar TIGER korpusa. Točnost svakog od algoritama smo provjerili analizom kratkog članka u svakom programu. Formula za izračun postotka pogreške glasi:

$$\text{Postotak pogreške} = \left(\frac{\text{Broj pogreški}}{\text{Ukupan broj riječi}}\right) \cdot 100$$

U svim alatima smo obradili odlomak iz članka koji je preuzet sa stranice njemačkog dnevnika *Zeit* te uspoređen s točnim rezultatima, koji uvažavaju sva pravila njemačke lingvistike. Odlomak iz njemačkog dnevnika se nalazi u privitku unutar poglavlja 5.5 Izvorni tekstovi pod oznakom (3).

U tablici 1 prikazana je točnost dvaju obilježivača vrste riječi za njemački jezik:

Obilježivač vrsta riječi	Ukupan broj riječi originalnog teksta	Broj pogreški	Postotak pogrešaka	Konačni rezultat točnosti
Pro3GresDE	**62**	**4**	**6,45%**	**93,55%**
Obilježivač TIGER korpusa	62	6	9,68%	90,32%

Tablica 1: Točnost obilježivača vrsta riječi Pro3GresDE i obilježivača uključenog unutar TIGER korpusa

Kao što primjećujemo Pro3GresDE obilježivač vrsta riječi čini neznatno manje pogrešaka u odnosu na čisti obilježivač vrsta riječi koji se nalazi unutar TIGER korpusa. Standardni problemi Pro3GresDE kao što su objekt u genitivu, imenski predikat u vokativu ili priložne funkcije nisu utjecali na dobivene rezultate zbog uspješnog rješavanja zadanih problema. Iako nije navedeno, Pro3GresDE je pokazao neznatne pogreške kod kontekstualnog raspoznavanja imena stranaka, koje je rastavljao te naknadno pokazivao kao zasebne vrste riječi, što u gramatičkom smislu nije krivo jer, *Labour-Partei* zapravo jest standardna imenica njemačkog jezika ili pak *Nazionalistischen Partei* koja je zapravo isto imenica s pridjevom u rodu broju i padežu, ali u ovom slučaju čisti naziv stranke kao takav. Još jedna pogreška jest u riječi

Negativität koju je ovdje predstavio kao osobno ime te joj stoga priključio oznaku za pripadajuću klasu, što svakako nije točno.

S druge strane osim navedenih pogreški koje je napravio Pro3GresDE, obilježivač vrsta riječi u TIGER korpusu je pokazao neznatno lošije rezultate. Pošto je Pro3GresDE baziran na TIGER korpusu te dorađen, riješen je problem označavanja osobnih imena s oznakom NN koja više služi za obilježavanje klase standardnih imenica (govorimo o imenicama koje ne simboliziraju naziv grada, osobnog imena, ustanove itd). Važno je napomenuti kako oba obilježivača koriste isti set oznaka, što uvelike olakšalo usporedbu.

3.2 Usporedba korjenovatelja

Kako smo uspoređivali obilježivače vrsta riječi isto tako ćemo uspoređivati korjenovatelje zasebno. Ovdje će se utvrditi njihova točnost na temelju tekstova (1) i (2) koji se nalaze u privitku pod poglavljem 5.5 izvorni tekstovi.

Uspoređuje se više korjenovatelja: UniNE korjenovatelj (lako korjenovanje i agresivno korjenovanje) u Perlu, Snowball korjenovatelj (unaprijeđena verzija), Text::German korjenovatelj u Perlu te CISTEM korjenovatelj na tekstu (2). Tekst (2) sam osobno sastavio s ciljem testiranja slabih točki svih alata. Tekst (2) sadrži 262 riječi koje nisu nužno napisane prema pravilima njemačke gramatike i lingvistike.

Korjenovatelj	Alat	Ukupan broj riječi	Broj grešaka	Postotak pogreške	Postotak ispravnosti
Snowball	Snowball Tartarus	262	46	17,56%	82,44%
CISTEM	**CISTEM**	**262**	**23**	**8,77%**	**91.23%**
UniNE (laki)	CPAN::Lingua:UniNE	262	56	21,37%	78,63%
UniNE (agresivni)	CPAN::Lingua:UniNE	262	50	19,08%	80,92%
Text::German	CPAN:Text:German	262	30	11,45%	88,55%

Tablica 2: Usporedba korjenovatelja Snowball, CISTEM, Txt::German te UniNE (jednostavno i agresivno korjeovanje)

Problem kod Snowball korjenovatelja leži u tome što korjenuje bilo koju riječ, nevažno o postojanju te riječi unutar njemačkog jezika, kao npr. *Tenot* ili *Kätzlein* ili *irg*. Sva tri primjera ne postoje u njemačkoj leksikografiji, te nisu riječi standardnog njemačkog jezika. Ne računajući predstavljanje svake imenice napisane velikim početnim slovom kao imenice s malim početnim slovom kao grešku, vidimo da svaki od korjenovatelja, pa tako i Snowball, ima problema s homografima ili istopisnicama. Također dolazi do nepotpunog rezanja nastavka i korjenovanja riječi u ponekim situacijama kao što su npr *aufweckt* ili *beliebt* (bilo je potrebno maknuti nastavak *t* u riječima, jer on ima ulogu ili nastavka 3. lica jednine ili nastavka participa perfekta glagola, dok u drugom primjeru riječ *bliebt* ne postoji).

CISTEM korjenovatelj, koji se pokazao najboljim u trenutnoj grupi, otklanja probleme koje je imao Snowball korjenovatelj unutar danog teksta, što znači da nema problema koje smo uočili na primjeru *beliebt* ili *aufweckt*. I dalje je nazočan problem homografa koji imaju dvojako značenje. Taj problem sam brojao kao jednu pogrešku, a ne kao dvije. CISTEM je precizniji od Snowball algoritma, iako kao i prethodni vrši supstituciju znaka *ß* sa *ss* unutar riječi kao npr. *Kuß* ili *daß* koji su po novijim standradima gramatike izbačeni iz pravopisa u određenim situacijama.

UniNE korjenovatelj (bilo riječ o jednostavnijem korjenovanju ili agresivnijem korjenovanju) se pokazao najnepreciznijim (naročito u jednostavnijem pristupu korjenovanja). Osim problema s homografima koji su imali svi korjenovatelji ovdje, te problema koje je imao Snowball, dodatno UniNE ne vrši korjenovanje nekih riječi do kraja, tj. ne uzima cijeli nastavak i primjenjuje pogrešna pravila korjenovanja. U njegovoj agresivnijoj verziji vidi se poboljšanje ali i dalje nedovoljno kvalitetni rezultati u odnosu na ostale.

Na posljetku testiran je Text::German korjenovatelj u CPAN modulu koji je imao slične probleme s homografima, zamjenom znaka *ß* sa *ss* u riječima *Kuß* i *daß*. U ponekim situacijama je prepoznao nepostojanje riječi u njemačkom jeziku kao npr. *Collagene* ili *Stubekcen* dok u drugima nije imao taj problem.

3.3 Usporedba lematizatora

U ovom dijelu uspoređujem dva lematizatora GermaLemma lematizator i SMOR lematizator. Alate sam opet testirao na istom tekstu (2) kao i korjenovatelje.

Lematizator	Alat	Ukupan broj riječi	Broj grešaka	Postotak pogreške	Postotak ispravnosti

GermaLemma	GermaLemma parzu	262	38	14,5%	85,5%
SMOR	**SMOR**	**262**	**15**	**5,73%**	**94,27%**

Tablica 3: Usporedba točnosti lematizatora SMOR i GermaLemma

SMOR lematizator se pokazao iznimno točnim. Sam po sebi eliminira sve imenice napisane malim početnim slovom, te riječi koje su izmišljene također odbacuje. Što se tiče dvoznačnosti, tj. homografije SMOR će definirati u nekim situacijama obje mogućnosti, tj. neće raspoznati koje se značenje trenutno traži već će ponuditi oba. U ponekim situacijama SMOR lematizator nije prepoznao da je riječ o homografiji te, i ako je prepoznao, jedna lema je predstavljala obje riječi uz ispis oba značenja. Također jedna od poznatijih grešaka jest prihvaćanje znaka *ß* u situacijama kada ne bi trebao biti prihvaćen zbog funkcije supstitucije primijenjenog unutar samog algoritma. Unutar testiranja nisam naišao na problem koji SMOR ima s ponekim nepravilnim glagolom, jer je za dani tekst svaki nepravilan glagol pravilno obradio. Osim navedenog pojavljuje se problem neprihvaćanja riječi *Chi, Chis* (njem. naziv za 22. slovo grčke abecede X-hi, te naziv slova na njemačkom jeziku u genitivu jednine).

GermaLemm je imao problem s prihvaćanjem imenica napisanih malim slovom. Važno je napomenuti kako sam instalirao *Pattern* paket koji povećava točnost lematizatora. Međutim za razliku od SMOR-a, ovaj lematizator je normalno korjenovao prethodni primjer riječi *Chi* i *Chis*. Najviše greški dolazi od homografa, tj. homografija te zamjena znaka *ß* sa *ss* pa radi toga prihvaća riječi *daß, Kuß* itd.

Uspoređujući oba alata SMOR se pokazao kao bolja mogućnost te je iz tog razloga jedan od proširenijih i poznatijih njemačkih lematizatora.

4 Zaključak

Nakon što sam obuhvatio 8 alata za računalnu analizu njemačkog jezika, rezultati pokazuju kako je još svakako potrebno doraditi alate kako bi pridonijeli točnosti obrade prirodnog jezika. U usporedbi od 4 korjenovatelja od najboljeg prema najlošijem su CISTEM korjenovatelj sa 91.23%, Text::German korjenovatelj sa 88,55%, nakon njega Snowball korjenovatelj sa 82,44% te na kraju UniNe korjenovatelj čija dva pristupa korjenovanju pridaju točnost u rasponu od 78,63% do 80,92%. Osim korjenovatelja uspoređena su i dva lematizatora, SMOR lematizator koji je pokazao točan u 94,27% slučajeva te GermaLemma, koji je bio točan u 85,5% slučajeva. Obilježivači vrsta riječi Pro3GresDE te onaj alat uključen unutar TIGER korpusa pokazali su respektabilne točnosti u iznosu 93,55% za Pro3GresDE te 90,32% za obilježivač vrsta riječi unutar TIGER korpusa. Prije samog zaključka važno je napomenuti kako osim navedenih opcija postoje i drugi alati čije točnosti nisam mogao isprobati zbog ograničenja licenci i pristupa. U trenutnoj analizi SMOR, CISTEM i Pro3GresDE su pokazali najbolje rezultate, dakako uz mogućnost korištenja TIGER korpusa. Svi alati uspješno analiziraju dane tekstove.

5 Privitak

5.1 Algoritam CISTEM korjenovatelja u Pythonu

```
#!/usr/bin/python3
"""
CISTEM Stemmer for German
This is the official Perl implementation of
the CISTEM stemmer.
It is based on the paper
Leonie Weißweiler, Alexander Fraser (2017).
Developing a Stemmer for German Based on a
Comparative Analysis of Publicly Available
Stemmers. In Proceedings of the German Society
for Computational Linguistics and Language
Technology (GSCL)
which can be read here:
http://www.cis.lmu.de/~weissweiler/cistem/
In the paper, we conducted an analysis of
publicly available stemmers, developed
two gold standards for German stemming and
evaluated the stemmers based on the
two gold standards. We then proposed the
stemmer implemented here and show
that it achieves slightly better f-measure
than the other stemmers and is
thrice as fast as the Snowball stemmer for
German while being about as fast as
most other stemmers.
"""
import re


stripge = re.compile(r"^ge(.{4,})")
replxx = re.compile(r"(.)\1")
replxxback = re.compile(r"(.)\*");
stripemr = re.compile(r"e[mr]$")
stripnd = re.compile(r"nd$")
stript = re.compile(r"t$")
stripesn = re.compile(r"[esn]$")


"""
This method takes the word to be stemmed and a
boolean specifiying if case-insensitive
stemming should be used and returns the
stemmed word. If only the word
```

```
is passed to the method or the second
parameter is 0, normal case-sensitive stemming
is used, if the second parameter is 1, case-
insensitive stemming is used.
Case sensitivity improves performance only if
words in the text may be incorrectly upper
case.
For all-lowercase and correctly cased text,
best performance is achieved by
using the case-sensitive version.
"""


def stem(word, case_insensitive = False):
    if len(word) == 0:
        return word


    upper = word[0].isupper()
    word = word.lower()


    word = word.replace("ü","u")
    word = word.replace("ö","o")
    word = word.replace("ä","a")
    word = word.replace("ß","ss")


    word = stripge.sub(r"\1", word)
    word = word.replace("sch","$")
    word = word.replace("ei","%")
    word = word.replace("ie","&")
    word = replxx.sub(r"\1*", word)


    while len(word) > 3:
        if len(word) > 5:
            (word, success) =
stripemr.subn("", word)
            if success != 0:
                continue


            (word, success) = stripnd.subn("",
word)
            if success != 0:
```

```
                continue


        if not upper or case_insensitive:
            (word, success) = stript.subn("",
word)
            if success != 0:
                continue


        (word, success) = stripesn.subn("",
word)
        if success != 0:
            continue
        else:
            break


    word = replxxback.sub(r"\1\1", word)
    word = word.replace("%","ei")
    word = word.replace("&","ie")
    word = word.replace("$","sch")


    return word


"""
This method works very similarly to stem. The
only difference is that in
addition to returning the stem, it also
returns the rest that was removed at
the end. To be able to return the stem
unchanged so the stem and the rest
can be concatenated to form the original word,
all subsitutions that altered
the stem in any other way than by removing
letters at the end were left out.
"""


def segment(word, case_insensitive = False):


    rest_length = 0
```

```
    if len(word) == 0:
        return ("", "")


    upper = word[0].isupper()
    word = word.lower()


    original = word[:]


    word = word.replace("sch","$")
    word = word.replace("ei","%")
    word = word.replace("ie","&")
    word = replxx.sub(r"\1*", word)


    while len(word) > 3:
        if len(word) > 5:
            (word, success) =
stripemr.subn("", word)
            if success != 0:
                rest_length += 2
                continue


            (word, success) = stripnd.subn("",
word)
            if success != 0:
                rest_length += 2
                continue


        if not upper or case_insensitive:
            (word, success) = stript.subn("",
word)
            if success != 0:
                rest_length += 1
                continue


        (word, success) = stripesn.subn("",
word)
        if success != 0:
            rest_length += 1
```

```
                continue
            else:
                break



        word = replxxback.sub(r"\1\1", word)
        word = word.replace("%","ei")
        word = word.replace("&","ie")
        word = word.replace("$","sch")


        if rest_length:
            rest = original[-rest_length:]
        else:
            rest = ""


        return (word,rest)
```

5.2 Algoritam u korjenovatelj za njemački jezik u Snowballu[20]

```
/*
    Extra rule for -nisse ending added 11 Dec 2009
*/

routines (
           prelude postlude
           mark_regions
           R1 R2
           standard_suffix
)

externals ( stem )
```

20 Preuzeto sa:*Snowball. Tartarus.*
http://snowball.tartarus.org/algorithms/german/stem_ISO_8859_1.sbl

```
integers ( p1 p2 x )

groupings ( v s_ending st_ending )

stringescapes {}

/* special characters (in ISO Latin I) */

stringdef a"   hex 'E4'
stringdef o"   hex 'F6'
stringdef u"   hex 'FC'
stringdef ss   hex 'DF'

define v 'aeiouy{a"}{o"}{u"}'

define s_ending  'bdfghklmnrt'
define st_ending s_ending - 'r'

define prelude as (

    test repeat (
        (
            ['{ss}'] <- 'ss'
        ) or next
    )

    repeat goto (
        v [('u'] v <- 'U') or
           ('y'] v <- 'Y')
    )
)

define mark_regions as (
```

```
    $p1 = limit
    $p2 = limit

    test(hop 3 setmark x)

    gopast v  gopast non-v  setmark p1
    try($p1 < x  $p1 = x)  // at least 3
    gopast v  gopast non-v  setmark p2

)

define postlude as repeat (

    [substring] among(
        'Y'    (<- 'y')
        'U'    (<- 'u')
        '{a"}' (<- 'a')
        '{o"}' (<- 'o')
        '{u"}' (<- 'u')
        ''     (next)
    )

)

backwardmode (

    define R1 as $p1 <= cursor
    define R2 as $p2 <= cursor

    define standard_suffix as (
        do (
            [substring] R1 among(
                'em' 'ern' 'er'
                (   delete
```

```
        )
        'e' 'en' 'es'
        (   delete
            try (['s'] 'nis' delete)
        )
        's'
        (   s_ending delete
        )
    )
)
do (
    [substring] R1 among(
        'en' 'er' 'est'
        (   delete
        )
        'st'
        (   st_ending hop 3 delete
        )
    )
)
do (
    [substring] R2 among(
        'end' 'ung'
        (   delete
            try (['ig'] not 'e' R2 delete)
        )
        'ig' 'ik' 'isch'
        (   not 'e' delete
        )
        'lich' 'heit'
        (   delete
            try (
                ['er' or 'en'] R1 delete
            )
```

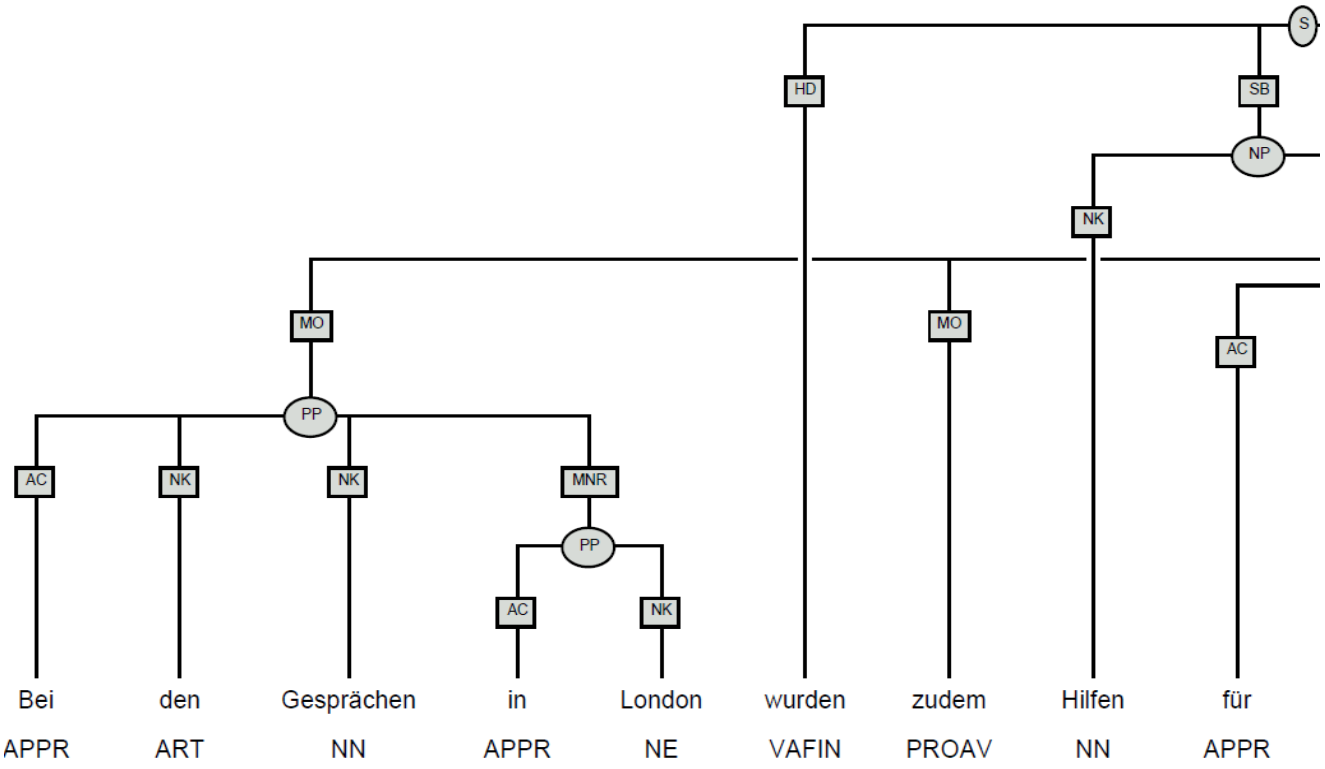
```
                )
                'keit'
                (   delete
                    try (
                        [substring] R2 among(
                            'lich' 'ig'
                            (   delete
                            )
                        )
                    )
                )
            )
        )
    )
)

define stem as (
    do prelude
    do mark_regions
    backwards
        do standard_suffix
    do postlude
)
```

5.3 Primjer rječnika sa korjenovanim oblicima pomoću Snowballa

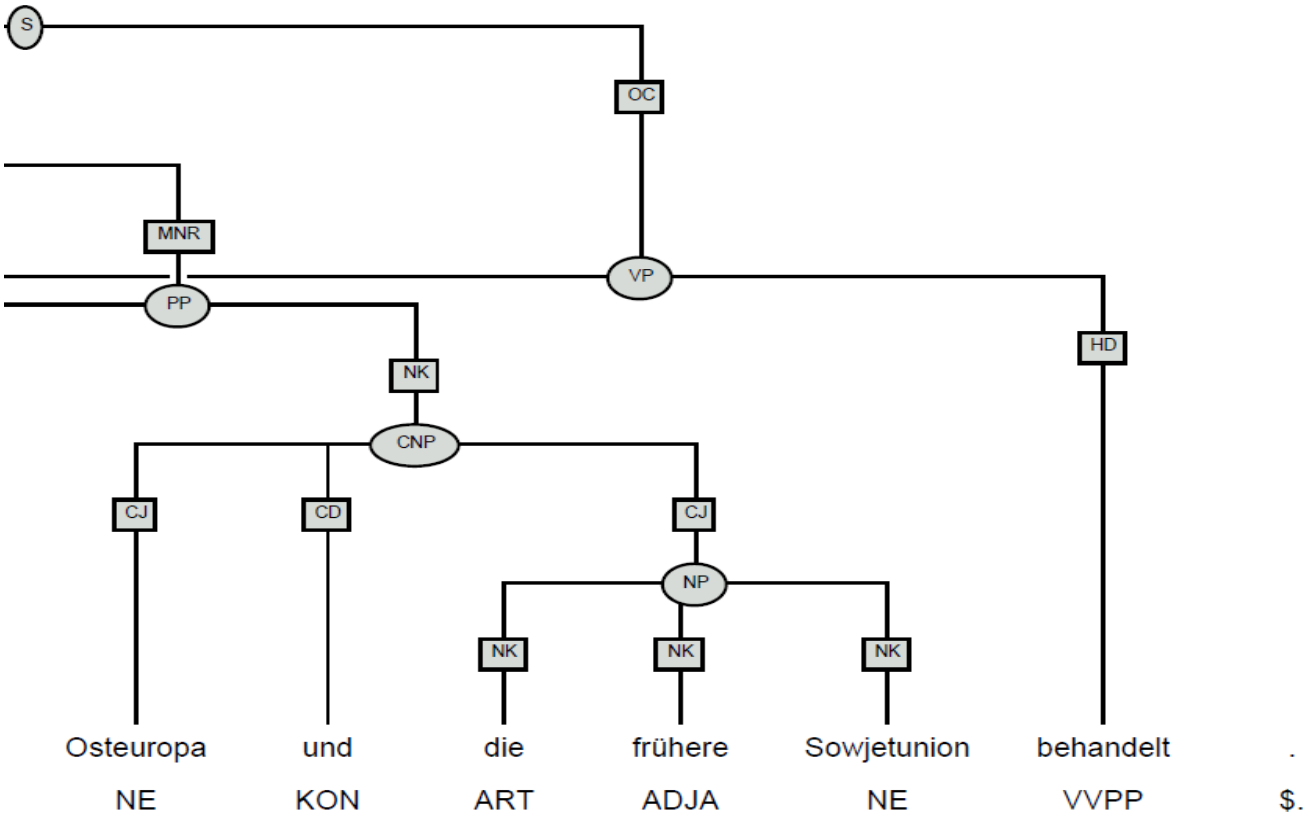
riječ	korijen	riječ	korijen
aufeinander	aufeinand	kategorie	kategori
aufeinanderbiss	aufeinanderbiss	kategorien	kategori
aufeinanderfolge	aufeinanderfolg	kategorisch	kategor
aufeinanderfolgen	aufeinanderfolg	kategorische	kategor
aufeinanderfolgend	aufeinanderfolg	kategorischen	kategor
Aufeinanderfolgenden	aufeinanderfolg	kategorischer	kategor
aufeinanderfolgende	aufeinanderfolg	kater	kat
aufeinanderfolgender	aufeinanderfol	katerliede	katerlied

aufeinanderfolgt	aufeinanderfolgt	katern	kat
aufeinanderfolgten	aufeinanderfolgt	katers	kat
aufeinanderschlügen	aufeinanderschlu	käthchen	kathch
aufenthalt	aufenthal	kathedrale	kathedral
aufenthalten	aufenthalt	kathinka	kathinka
aufenthaltes	aufenthalt	katholik	kathol
auferlegen	auferleg	katholische	katho
auferlegt	auferlegt	katholischen	katho
auferlegten	auferlegt	katholischer	katho
auferstand	auferstand	kattun	kattun
auferstanden	auferstand	kattunhalstücher	kattunhalstuc
auferstehen	aufersteh	katz	katz
aufersteht	aufersteht	kätzchen	katzch
auferstehung	aufersteh	kätzchens	katzch
auferstünde	auferstund	katzenschmer	katzenschm
auferwecken	auferweck	katzensprung	katzenspr
auferweckt	auferweckt	katzenwürde	katzenwurd
auferzogen	auferzog	kätzin	katzin
aufessen	aufess	kätzlein	katzlein
auffa	auffa	katzmann	katzmann
auffallen	auffall	kauen	kau
auffallend	auffall	kauerte	kauert
auffallenden	auffall	kauf	kauf
auffallender	auffall	kaufe	kauf
auffällig	auffall	kaufen	kauf
auffälligen	auffall	käufer	kauf
auffälliges	auffall	kauffahrer	kauffahr
auffassen	auffass	kaufherr	kaufherr
auffasst	auffasst	kaufleute	kaufleut
auffaßt	auffasst	käuflich	kauflich
auffassung	auffass		
auffassungsvermögen	auffassungsvermog		

5.4 Primjer stabla TIGER korpusa



Slika 11: Lijeva grana stabla (Telljohann, Hinrichs, Kübler. 2004)



Slika 12: Desna grana stabla (Telljohann, Hinrichs, Kübler. 2004)

5.5 Izvorni tekstovi

(1)[21]

Maltas Ministerpräsident Joseph Muscat hat seine Partei zum Gewinner der vorgezogenen Parlamentswahlen erklärt. Inoffizielle Ergebnisse deuten auf einen "beträchtlichen" Sieg für seine Labour-Partei hin, sagte Muscat im nationalen Rundfunk. Muscat zufolge hätten sich die Wähler damit für "die positive Kampagne" seiner Partei und nicht für die "Negativität und Bitterkeit der Nationalistischen Partei" entschieden. Sein Rivale Simon Busuttil hatte bereits seine Niederlage eingeräumt.

(2)

Adlers neutrales adler Adler Neutrales Verlierer verlirer machen Zustand Beispiel Maus maus zerstören Zerstörung Küsse Kuss Kuß Störsender verlies Verlies stöhnen singen singt beliebt Kuß Küsse beliebster schawache Mutter mutter vater Vater aufeinander aufeinanderfolgenden aufwecken kätzin auffallen auffallenden auffällig katz kattun kätzchen katzen kätzin kategorischen käuflich kaufherr käufer kauffahrer Stuck Eisbär Eisbar Stück Häuser Haus Apfel Äpfel Tempel tempel mach Mach Sterben sterben starb davor deutschlands Deutschlands schönes schön schon Schon Kanton kanton Fahrrad Austauschstudenten hüten dass daß das sich baut Politiker politiker abschließender Urteilsgabe urteilsgabe Aale Ahle Aar Ahr modern modern Gliedersatz Heorin heroin Konstanz Stubecken Staubecken aufeinanderfolgenser aufeinanderfolgt aufenhalt aufenhalten aufenthalten Aufenhalten Aufenthalten katholik Katholik kattun katholische katholischen katern katedrale Kathedrale auferlegen Aufeinanderfolgenden aufeinanderfolgenden kategorie Kategorie Kategorien Kategorien übersetzen übersetzen Übersetzen Wachstube wachstube Wachstube Versendung umgehen umgehen Umgehen Tenot Tenor Tenor Heorin heroin Heroin Abort Abort abort Band Band Back Boot Back Boot boot back Collagen Collagen Collagens Collagene des der der des Aborts Abortes Aborte Aborten aborten aufweckt auferzogen aufessen auffa auffallen kauen kauerte kauf kaufe kaufen käufer Käufer käuflich kätzchen kätzchens kathedral Kathedral kathe kathe Kathe katers Kätzin Kätzlein kauerte abtritt Abtritt Abtri abtri Häusl Haus haus hausen Hausen Häuser hasen Hasen Hase hassen Hassen irgendwann wann Wann irgend irg worann woran Wirtsstube Sickergrube sickergrube dazugehörige Abortschlüssel Abortspülung Abortbrille ab

[21] Preuzeto sa: https://www.zeit.de/politik/ausland/2017-06/malta-joseph-muscat-labour-neuwahlen-panama-papers

an auf aufeinander einander Abortgrube Abortpapier Chi Chis chis durchtrennen trennen durch setzen durchsetzen Eis Eises Eise Eis Reime Hörbeispiele siet vergleiche vergleich Vergleich Speiseeis Speise spaß spass niederdeutsch nieder deutsch österreichisch Österreich See See Wasser Milcheis Vanileeis Vanile Essener essener fluchten Fluchten Fluchtung Lexika Lexiken Islam Islams

5.6 Rezultati obilježivača vrsta riječi Pro3GresDE i TIGER korpusa

5.6.1 Rezultati Pro3GresDE obilježivača

Pogreške su označene zelenom bojom:

Maltas NE Ministerpräsident NN Joseph NE Muscat NE hat VAFIN seine PPOSAT Partei NN zum APPRART Gewinner NN der ART vorgezogenen ADJA Parlamentswahlen NN erklärt VVPP . $. Inoffizielle ADJA Ergebnisse NE deuten VVFIN auf APPR einen ART " $(beträchtlichen ADJA " $(Sieg NN für APPR seine PPOSAT Labour-Partei NN hin PTKVZ , $, sagte VVFIN Muscat NE im APPRART nationalen ADJA Rundfunk NN . $. Muscat NE zufolge APPO hätten VAFIN sich PRF die ART Wähler NN damit PROP für APPR " $(die ART positive ADJA Kampagne NN " $(seiner PPOSAT Partei NN und KON nicht PTKNEG PTKNEG für APPR die ART " $(Negativität NN und KON Bitterkeit NN die ART Nationalistischen ADJA Partei NN " $(entscheiden VVPP Sein PPOSAT Rivale NN Simon NE Busuttil NE hatte VAFIN bereits ADV seine ART Niederlage NN eingeräumt VVPP . $.

5.6.2 Rezultati obilježivača vrsta riječi TIGER korpusa

Pogreške su označene zelenom bojom:

Maltas NE Ministerpräsident NN Joseph NE Muscat NE hat VAFIN seine PPOSAT Partei NN zum APPRART Gewinner NN der ART vorgezogenen ADJA Parlamentswahlen NN erklärt VVPP . $. Inoffizielle ADJA Ergebnisse NN deuten VVFIN auf APPR einen ART " $(beträchtlichen ADJA " $(Sieg NN für APPR seine PPOSAT Labour-Partei NN hin PTKVZ , $, sagte VVFIN Muscat NN im APPRART nationalen ADJA Rundfunk NN . $. Muscat NN zufolge APPO hätten VAFIN sich PRF die ART Wähler NN damit PROP für APPR " $(die ART positive ADJA Kampagne NN " $(seiner PPOSAT Partei NN und KON nicht PTKNEG PTKNEG für APPR die ART " $(Negativität NE und KON Bitterkeit NN die ART Nationalistischen ADJA Partei NN " $(entscheiden VVPP Sein PPOSAT Rivale NN Simon NE Busuttil NE hatte VAFIN bereits ADV seine ART Niederlage NN eingeräumt VVPP . $.

5.7 Rezultati korjenovatelja

5.7.1 Rezultati Snowball korjenovatelja

Pogreške algoritma su značene zelenom bojom:

Adlers neutrales adler Adler Neutrales Verlierer verlirer machen Zustand Beispiel Maus maus zerstören Zerstörung Küsse Kuss Kuß Störsender verlies Verlies stöhnen singen singt beliebt Kuß Küsse beliebster schawache Mutter mutter vater Vater aufeinander aufeinanderfolgenden aufwecken kätzin auffallen auffallenden auffällig katz kattun kätzchen katzen kätzin kategorischen käuflich kaufherr käufer kauffahrer Stuck Eisbär Eisbar Stück Häuser Haus Apfel Äpfel Tempel tempel mach Mach Sterben sterben starb davor deutschlands Deutschlands schönes schön schon Schon Kanton kanton Fahrrad Austauschstudenten hüten dass daß das sich baut Politiker politiker abschließender Urteilsgabe urteilsgabe Aale Ahle Aar Ahr modern modern Gliedersatz Heorin heroin Konstanz Stubecken Staubecken aufeinanderfolgenser aufeinanderfolgt aufenhalt aufenhalten aufenthalten Aufenhalten Aufenthalten katholik Katholik kattun katholische katholischen katern katedrale Kathedrale auferlegen Aufeinanderfolgenden aufeinanderfolgenden kategorie Kategorie Kategorien Kategorien übersetzen übersetzen Übersetzen Wachstube wachstube Wachstube Versendung umgehen umgehen Umgehen Tenot Tenor Tenor Heorin heroin Heroin Abort Abort abort Band Band Back Boot Back Boot boot back Collagen Collagen Collagens Collagene des der der des Aborts Abortes Aborte Aborten aborten aufweckt auferzogen aufessen auffa auffallen kauen kauerte kauf kaufe kaufen käufer Käufer käuflich kätzchen kätzchens kathedral Kathedral kathe kathe Kathe katers Kätzin Kätzlein kauerte abtritt Abtritt Abtri abtri Häusl Haus haus hausen Hausen Häuser hasen Hasen Hase hassen Hassen irgendwann wann Wann irgend irg worann woran Wirtsstube Sickergrube sickergrube dazugehörige Abortschlüssel Abortspülung Abortbrille ab an auf aufeinander einander Abortgrube Abortpapier Chi Chis chis durchtrennen trennen durch setzen durchsetzen Eis Eises Eise Eis Reime Hörbeispiele siet vergleiche vergleich Vergleich Speiseeis Speise spaß spass niederdeutsch nieder deutsch österreichisch Österreich See See Wasser Milcheis Vanileeis Vanile Essener essener fluchten Fluchten Fluchtung Lexika Lexiken Islam Islams

5.7.2 Rezultati CISTEM korjenovatelja

Pogreške algoritma su značene zelenom bojom:

Adlers neutrales adler Adler Neutrales Verlierer verlirer machen Zustand Beispiel Maus maus zerstören Zerstörung Küsse Kuss Kuß Störsender verlies Verlies stöhnen singen singt beliebt

Kuß Küsse beliebster schawache Mutter mutter vater Vater aufeinander aufeinanderfolgenden aufwecken kätzin auffallen auffallenden auffällig katz kattun kätzchen katzen kätzin kategorischen käuflich kaufherr käufer kauffahrer Stuck Eisbär Eisbar Stück Häuser Haus Apfel Äpfel Tempel tempel mach Mach Sterben sterben starb davor deutschlands Deutschlands schönes schön schon Schon Kanton kanton Fahrrad Austauschstudenten hüten dass daß das sich baut Politiker politiker abschließender Urteilsgabe urteilsgabe Aale Ahle Aar Ahr modern modern Gliedersatz Heorin heroin Konstanz Stubecken Staubecken aufeinanderfolgenser aufeinanderfolgt aufenhalt aufenhalten aufenthalten Aufenhalten Aufenthalten katholik Katholik kattun katholische katholischen katern katedrale Kathedrale auferlegen Aufeinanderfolgenden aufeinanderfolgenden kategorie Kategorie Kategorien Kategorien übersetzen übersetzen Übersetzen Wachstube wachstube Wachstube Versendung umgehen umgehen Umgehen Tenot Tenor Tenor Heorin heroin Heroin Abort Abort abort Band Band Back Boot Back Boot boot back Collagen Collagen Collagens Collagene des der der des Aborts Abortes Aborte Aborten aborten aufweckt auferzogen aufessen auffa auffallen kauen kauerte kauf kaufe kaufen käufer Käufer käuflich kätzchen kätzchens kathedral Kathedral kathe kathe Kathe katers Kätzin Kätzlein kauerte abtritt Abtritt Abtri abtri Häusl Haus haus hausen Hausen Häuser hasen Hasen Hase hassen Hassen irgendwann wann Wann irgend irg worann woran Wirtsstube Sickergrube sickergrube dazugehörige Abortschlüssel Abortspülung Abortbrille ab an auf aufeinander einander Abortgrube Abortpapier Chi Chis chis durchtrennen trennen durch setzen durchsetzen Eis Eises Eise Eis Reime Hörbeispiele siet vergleiche vergleich Vergleich Speiseeis Speise spaß spass niederdeutsch nieder deutsch österreichisch Österreich See See Wasser Milcheis Vanileeis Vanile Essener essener fluchten Fluchten Fluchtung Lexika Lexiken Islam Islams

5.7.3 Rezultati Text::German korjenovatelja

Pogreške algoritma su značene zelenom bojom:

Adlers neutrales adler Adler Neutrales Verlierer verlirer machen Zustand Beispiel Maus maus zerstören Zerstörung Küsse Kuss Kuß Störsender verlies Verlies stöhnen singen singt beliebt Kuß Küsse beliebster schawache Mutter mutter vater Vater aufeinander aufeinanderfolgenden aufwecken kätzin auffallen auffallenden auffällig katz kattun kätzchen katzen kätzin kategorischen käuflich kaufherr käufer kauffahrer Stuck Eisbär Eisbar Stück Häuser Haus Apfel Äpfel Tempel tempel mach Mach Sterben sterben starb davor deutschlands Deutschlands schönes schön schon Schon Kanton kanton Fahrrad Austauschstudenten hüten dass daß das sich baut Politiker politiker abschließender Urteilsgabe urteilsgabe Aale Ahle Aar Ahr modern

modern Gliedersatz Heorin heroin Konstanz Stubecken Staubecken aufeinanderfolgenser aufeinanderfolgt aufenhalt aufenhalten aufenthalten Aufenhalten Aufenthalten katholik Katholik kattun katholische katholischen katern katedrale Kathedrale auferlegen Aufeinanderfolgenden aufeinanderfolgenden kategorie Kategorie Kategorien Kategorien übersetzen übersetzen Übersetzen Wachstube wachstube Wachstube Versendung umgehen umgehen Umgehen Tenot Tenor Tenor Heorin heroin Heroin Abort Abort abort Band Band Back Boot Back Boot boot back Collagen Collagen Collagens Collagene des der der des Aborts Abortes Aborte Aborten aborten aufweckt auferzogen aufessen auffa auffallen kauen kauerte kauf kaufe kaufen käufer Käufer käuflich kätzchen kätzchens kathedral Kathedral kathe kathe Kathe katers Kätzin Kätzlein kauerte abtritt Abtritt Abtri abtri Häusl Haus haus hausen Hausen Häuser hasen Hasen Hase hassen Hassen irgendwann wann Wann irgend irg worann woran Wirtsstube Sickergrube sickergrube dazugehörige Abortschlüssel Abortspülung Abortbrille ab an auf aufeinander einander Abortgrube Abortpapier Chi Chis chis durchtrennen trennen durch setzen durchsetzen Eis Eises Eise Eis Reime Hörbeispiele siet vergleiche vergleich Vergleich Speiseeis Speise spaß spass niederdeutsch nieder deutsch österreichisch Österreich See See Wasser Milcheis Vanileeis Vanile Essener essener fluchten Fluchten Fluchtung Lexika Lexiken Islam Islams

5.7.4 Rezultati UniNE korjenovatelja

Pogreške algoritma su značene zelenom bojom:

5.7.4.1 Jednostavno korjenovanje

Adlers neutrales adler Adler Neutrales Verlierer verlirer machen Zustand Beispiel Maus maus zerstören Zerstörung Küsse Kuss Kuß Störsender verlies Verlies stöhnen singen singt beliebt Kuß Küsse beliebster schawache Mutter mutter vater Vater aufeinander aufeinanderfolgenden aufwecken kätzin auffallen auffallenden auffällig katz kattun kätzchen katzen kätzin kategorischen käuflich kaufherr käufer kauffahrer Stuck Eisbär Eisbar Stück Häuser Haus Apfel Äpfel Tempel tempel mach Mach Sterben sterben starb davor deutschlands Deutschlands schönes schön schon Schon Kanton kanton Fahrrad Austauschstudenten hüten dass daß das sich baut Politiker politiker abschließender Urteilsgabe urteilsgabe Aale Ahle Aar Ahr modern modern Gliedersatz Heorin heroin Konstanz Stubecken Staubecken aufeinanderfolgenser aufeinanderfolgt aufenhalt aufenhalten aufenthalten Aufenhalten Aufenthalten katholik Katholik kattun katholische katholischen katern katedrale Kathedrale auferlegen Aufeinanderfolgenden aufeinanderfolgenden kategorie Kategorie Kategorien Kategorien übersetzen übersetzen Übersetzen Wachstube wachstube Wachstube Versendung umgehen

umgehen Umgehen Tenot Tenor Tenor Heorin heroin Heroin Abort Abort abort Band Band Back Boot Back Boot boot back Collagen Collagen Collagens Collagene des der der des Aborts Abortes Aborte Aborten aborten aufweckt auferzogen aufessen auffa auffallen kauen kauerte kauf kaufe kaufen käufer Käufer käuflich kätzchen kätzchens kathedral Kathedral kathe kathe Kathe katers Kätzin Kätzlein kauerte abtritt Abtritt Abtri abtri Häusl Haus haus hausen Hausen Häuser hasen Hasen Hase hassen Hassen irgendwann wann Wann irgend irg worann woran Wirtsstube Sickergrube sickergrube dazugehörige Abortschlüssel Abortspülung Abortbrille ab an auf aufeinander einander Abortgrube Abortpapier Chi Chis chis durchtrennen trennen durch setzen durchsetzen Eis Eises Eise Eis Reime Hörbeispiele siet vergleiche vergleich Vergleich Speiseeis Speise spaß spass niederdeutsch nieder deutsch österreichisch Österreich See See Wasser Milcheis Vanileeis Vanile Essener essener fluchten Fluchten Fluchtung Lexika Lexiken Islam Islams

5.7.4.2 Agresivno korjenovanje

Adlers neutrales adler Adler Neutrales Verlierer verlirer machen Zustand Beispiel Maus maus zerstören Zerstörung Küsse Kuss Kuß Störsender verlies Verlies stöhnen singen singt beliebt Kuß Küsse beliebster schawache Mutter mutter vater Vater aufeinander aufeinanderfolgenden aufwecken kätzin auffallen auffallenden auffällig katz kattun kätzchen katzen kätzin kategorischen käuflich kaufherr käufer kauffahrer Stuck Eisbär Eisbar Stück Häuser Haus Apfel Äpfel Tempel tempel mach Mach Sterben sterben starb davor deutschlands Deutschlands schönes schön schon Schon Kanton kanton Fahrrad Austauschstudenten hüten dass daß das sich baut Politiker politiker abschließender Urteilsgabe urteilsgabe Aale Ahle Aar Ahr modern modern Gliedersatz Heorin heroin Konstanz Stubecken Staubecken aufeinanderfolgenser aufeinanderfolgt aufenhalt aufenhalten aufenthalten Aufenhalten Aufenthalten katholik Katholik kattun katholische katholischen katern katedrale Kathedrale auferlegen Aufeinanderfolgenden aufeinanderfolgenden kategorie Kategorie Kategorien Kategorien übersetzen übersetzen Übersetzen Wachstube wachstube Wachstube Versendung umgehen umgehen Umgehen Tenot Tenor Tenor Heorin heroin Heroin Abort Abort abort Band Band Back Boot Back Boot boot back Collagen Collagen Collagens Collagene des der der des Aborts Abortes Aborte Aborten aborten aufweckt auferzogen aufessen auffa auffallen kauen kauerte kauf kaufe kaufen käufer Käufer käuflich kätzchen kätzchens kathedral Kathedral kathe kathe Kathe katers Kätzin Kätzlein kauerte abtritt Abtritt Abtri abtri Häusl Haus haus hausen Hausen Häuser hasen Hasen Hase hassen Hassen irgendwann wann Wann irgend irg worann woran Wirtsstube Sickergrube sickergrube dazugehörige Abortschlüssel Abortspülung Abortbrille ab

an auf aufeinander einander Abortgrube Abortpapier Chi Chis chis durchtrennen trennen durch setzen durchsetzen Eis Eises Eise Eis Reime Hörbeispiele siet vergleiche vergleich Vergleich Speiseeis Speise spaß spass niederdeutsch nieder deutsch österreichisch Österreich See See Wasser Milcheis Vanileeis Vanile Essener essener fluchten Fluchten Fluchtung Lexika Lexiken Islam Islams

5.8 Rezultati lematizatora

5.8.1 SMOR lematizator

Pogreške algoritma su značene zelenom bojom:

Adlers neutrales adler Adler Neutrales Verlierer verlirer machen Zustand Beispiel Maus maus zerstören Zerstörung Küsse Kuss Kuß Störsender verlies Verlies stöhnen singen singt beliebt Kuß Küsse beliebster schawache Mutter mutter vater Vater aufeinander aufeinanderfolgenden aufwecken kätzin auffallen auffallenden auffällig katz kattun kätzchen katzen kätzin kategorischen käuflich kaufherr käufer kauffahrer Stuck Eisbär Eisbar Stück Häuser Haus Apfel Äpfel Tempel tempel mach Mach Sterben sterben starb davor deutschlands Deutschlands schönes schön schon Schon Kanton kanton Fahrrad Austauschstudenten hüten dass daß das sich baut Politiker politiker abschließender Urteilsgabe urteilsgabe Aale Ahle Aar Ahr modern modern Gliedersatz Heorin heroin Konstanz Stubecken Staubecken aufeinanderfolgenser aufeinanderfolgt aufenhalt aufenhalten aufenthalten Aufenhalten Aufenthalten katholik Katholik kattun katholische katholischen katern katedrale Kathedrale auferlegen Aufeinanderfolgenden aufeinanderfolgenden kategorie Kategorie Kategorien Kategorien übersetzen übersetzen Übersetzen Wachstube wachstube Wachstube Versendung umgehen umgehen Umgehen Tenot Tenor Tenor Heorin heroin Heroin Abort Abort abort Band Band Back Boot Back Boot boot back Collagen Collagen Collagens Collagene des der der des Aborts Abortes Aborte Aborten aborten aufweckt auferzogen aufessen auffa auffallen kauen kauerte kauf kaufe kaufen käufer Käufer käuflich kätzchen kätzchens kathedral Kathedral kathe kathe Kathe katers Kätzin Kätzlein kauerte abtritt Abtritt Abtri abtri Häusl Haus haus hausen Hausen Häuser hasen Hasen Hase hassen Hassen irgendwann wann Wann irgend irg worann woran Wirtsstube Sickergrube sickergrube dazugehörige Abortschlüssel Abortspülung Abortbrille ab an auf aufeinander einander Abortgrube Abortpapier Chi Chis chis durchtrennen trennen durch setzen durchsetzen Eis Eises Eise Eis Reime Hörbeispiele siet vergleiche vergleich Vergleich Speiseeis Speise spaß spass niederdeutsch nieder deutsch österreichisch Österreich See See

Wasser Milcheis Vanileeis Vanile Essener essener fluchten Fluchten Fluchtung Lexika Lexiken Islam Islams

5.8.2 GermaLemma lematizator

Pogreške algoritma su značene zelenom bojom:

Adlers neutrales adler Adler Neutrales Verlierer verlirer machen Zustand Beispiel Maus maus zerstören Zerstörung Küsse Kuss Kuß Störsender verlies Verlies stöhnen singen singt beliebt Kuß Küsse beliebster schawache Mutter mutter vater Vater aufeinander aufeinanderfolgenden aufwecken kätzin auffallen auffallenden auffällig katz kattun kätzchen katzen kätzin kategorischen käuflich kaufherr käufer kauffahrer Stuck Eisbär Eisbar Stück Häuser Haus Apfel Äpfel Tempel tempel mach Mach Sterben sterben starb davor deutschlands Deutschlands schönes schön schon Schon Kanton kanton Fahrrad Austauschstudenten hüten dass daß das sich baut Politiker politiker abschließender Urteilsgabe urteilsgabe Aale Ahle Aar Ahr modern modern Gliedersatz Heorin heroin Konstanz Stubecken Staubecken aufeinanderfolgenser aufeinanderfolgt aufenhalt aufenhalten aufenthalten Aufenhalten Aufenthalten katholik Katholik kattun katholische katholischen katern katedrale Kathedrale auferlegen Aufeinanderfolgenden aufeinanderfolgenden kategorie Kategorie Kategorien Kategorien übersetzen übersetzen Übersetzen Wachstube wachstube Wachstube Versendung umgehen umgehen Umgehen Tenot Tenor Tenor Heorin heroin Heroin Abort Abort abort Band Band Back Boot Back Boot boot back Collagen Collagen Collagens Collagene des der der des Aborts Abortes Aborte Aborten aborten aufweckt auferzogen aufessen auffa auffallen kauen kauerte kauf kaufe kaufen käufer Käufer käuflich kätzchen kätzchens kathedral Kathedral kathe kathe Kathe katers Kätzin Kätzlein kauerte abtritt Abtritt Abtri abtri Häusl Haus haus hausen Hausen Häuser hasen Hasen Hase hassen Hassen irgendwann wann Wann irgend irg worann woran Wirtsstube Sickergrube sickergrube dazugehörige Abortschlüssel Abortspülung Abortbrille ab an auf aufeinander einander Abortgrube Abortpapier Chi Chis chis durchtrennen trennen durch setzen durchsetzen Eis Eises Eise Eis Reime Hörbeispiele siet vergleiche vergleich Vergleich Speiseeis Speise spaß spass niederdeutsch nieder deutsch österreichisch Österreich See See Wasser Milcheis Vanileeis Vanile Essener essener fluchten Fluchten Fluchtung Lexika Lexiken Islam Islams

5.9 Evaluacijski set 1 CISTEM korjenovatelja[22]

Supermann ermannest Restpostens schraffiertest restloses mannschaftlicher
entmannten mannhaftst Klassizismus bemannst bemannten ermannet restlos
ermannst mannhafteste bemannest bemannend mannhaftem mannhaften Restsummen
Supermänner schraffiert ermanntest Immobilienhändler restliche mannhafterer
Supermannes schraffierst Schraffierung entmanntet restlosen entmanntest
mannhaftes Mannschaft schraffierte bemannt ermannt Supermännern
Immobilienhändlern entmanne Männlein restlose schraffiertet mannhafte
ermanne Männchens bemannen entmannst mannschaftliches mannhaftest
mannhaftste mannhafteren mannhaftestem mannhafter entmannest schraffieren
Supermanne mannhaftester mannhaftster Neoklassizismen entmannte Supermanns
schraffierten restloser Restposten restlichen mannhaftesten mannhafterem
ermannen bemannet mannhaftstes mannhaftstem Neoklassizismus Klassizismen
mannschaftlichen Männchen mannschaftlichem ermannte Männleins entmannend
bemannte restlichem ermannend mannschaftlich entmannt bemanntet
mannhaftestes entmannen restlich Mannhaftigkeit schraffiere Restsumme
restliches mannhaftere Immobilienhändlers bemanntest Mannhaftigkeiten
ermannten ermanntet restlosem schraffieret Mannschaften mannhafteres
mannhaftsten schraffierend mannschaftliche restlicher mannhaft
Schraffierungen bemanne entmannet schraffierest
Abakus
Abandon Abandons
Abasien Abasie
Abdomens Abdomina Abdomen
Aberrationen Aberration
Abiogenese
Abitur Abituren Abiture Abiturs
Ablationen Ablation
Ablativen Ablative Ablativ Ablativs
Ablativus Ablativi
Abonnemente Abonnementen Abonnements Abonnement
Abrakadabra Abrakadabras
Abrasion Abrasionen
Absorbens Absorbenzien Absorbentia
Abszessen Abszesses Abszeß Abszesse
Abszisse Abszissen
Abyssus
Accessoires Accessoire
Acetons Aceton
Achensees Achenseen Achensee
Achillesfersen Achillesferse
Achillessehnen Achillessehne
Achromatopsie Achromatopsien
Achänen Achäne
Acres Acre
Adagios Adagio
Adamsapfels Adamsäpfeln Adamsäpfel Adamsapfel
Adamskostümen Adamskostüm Adamskostüms Adamskostüme
Addendum Addendums Addenda
Adenome Adenoms Adenom Adenomen
Adepten Adept
Adhäsionen Adhäsion
Adjektiven Adjektiv Adjektivs Adjektives Adjektive
Adjunkten Adjunkt

---

22 Preuzeto sa:

https://github.com/LeonieWeissweiler/CISTEM/blob/master/gold_standards/goldstandard2.txt

Adjutant Adjutanten
Adjutum Adjuten Adjutums
Admiralen Admiral Admirale Admirals
Adnexen Adnexe Adnexes Adnex
Adrenalins Adrenalin
Adria
Adventes Advents Advent Adventen Advente
Adverbs Adverbien Adverb
Advokaten Advokat
Aerobiont Aerobionten
Aeroflot
Aerolithe Aerolith Aeroliths Aerolithen
Aerologie
Aeroplanes Aeroplans Aeroplan Aeroplanen Aeroplane
Affekte Affekt Affekts Affektes Affekten
Affidavits Affidavit
Affix Affixes Affixe Affixen
Affodilles Affodills Affodill Affodillen Affodille
Affrikata Affrikaten
Affrikaten Affrikate
Affronts Affront
Affären Affäre
Afghane Afghanen
Afghanistans Afghanistan
Afrika Afrikas
Afrikaander Afrikaandern Afrikaanders
Afrikaans
Afrikanders Afrikandern Afrikander
Afrikanern Afrikaner Afrikaners
Agenda Agenden
Agens Agenzien
Agent Agenten
Aggression Aggressionen
Agnosien Agnosie
Agonisten Agonist
Agraffen Agraffe
Agraphien Agraphie
Agronom Agronomen
Aiden Aide
Aigrette Aigretten
Ainu Ainus
Airbus Airbussen Airbusse Airbusses
Akademie Akademien
Akanthites Akanthiten Akanthit Akanthits Akanthite
Akanthus
Akanthusblatte Akanthusblättern Akanthusblatt Akanthusblatts
Akanthusblattes Akanthusblätter
Akaroidharze Akaroidharzes Akaroidharz Akaroidharzen
Akkoladen Akkolade
Akkorden Akkords Akkorde Akkordes Akkord
Akkusativ Akkusative Akkusativen Akkusativs
Akontos Akonto
Akrobat Akrobaten
Akroleins Akrolein
Akromegalien Akromegalie
Akronyme Akronym Akronyms Akronymen
Akrostichon Akrostichen Akrosticha Akrostichons
Akroterie Akroterien
Akroterions Akroterion Akroterien
Akteur Akteuren Akteure Akteurs
Aktinien Aktinie

Aktivitas Aktivitates
Aktricen Aktrice
Aktuars Aktuare Aktuaren Aktuar

5.10 Evaluacijski set 2 CISTEM korjenovatelja

A
Aale Aal Aalen Aales Aals
Aars Aar Aaren Aare Aares
Aases Aas Aasen Äsern Äser Aase
Aasgeier Aasgeiern Aasgeiers
Abakus
Abandons Abandon
Abart Abarten
Abartung Abartungen
Abasie Abasien
Abbau Abbaue Abbauten Abbaus Abbauen Abbaues
Abbaufeldern Abbaufelder Abbaufelde Abbaufelds Abbaufeldes Abbaufeld
Abbaugerechtigkeit
Abbaurechts Abbaurecht Abbaurechtes Abbaurechten Abbaurechte
Abberufung Abberufungen
Abbestellungen Abbestellung
Abbilder Abbildes Abbild Abbilde Abbilds Abbildern
Abbildungen Abbildung
Abbitten Abbitte
Abblendlichte Abblendlichtes Abblendlichts Abblendlicht
Abbrand Abbrands Abbrande Abbrandes Abbränden Abbrände
Abbreviationen Abbreviation
Abbreviatur Abbreviaturen
Abbruches Abbrüchen Abbruchs Abbruche Abbruch Abbrüche
Abbrucharbeiten Abbrucharbeit
Abbrändler Abbrändlern Abbrändlers
Abbröckelung Abbröckelungen
Abdachungen Abdachung
Abdampfes Abdampf Abdämpfe Abdämpfen Abdampfs Abdampfe
Abdampfwärme
Abdankung Abdankungen
Abdecker Abdeckern Abdeckers
Abdeckereien Abdeckerei
Abdeckung Abdeckungen
Abderiten Abderit
Abdichtungen Abdichtung
Abdikationen Abdikation
Abdomens Abdomen Abdomina
Abdriften Abdrift
Abdrucks Abdruck Abdruckes Abdrucke Abdrucken
Abduktionen Abduktion
Abduktoren Abduktors Abduktor
Abende Abenden Abends Abend
Abendanzuges Abendanzügen Abendanzugs Abendanzuge Abendanzüge Abendanzug
Abendbrote Abendbroten Abendbrots Abendbrotes Abendbrot
Abenddämmerungen Abenddämmerung
Abendessen Abendessens
Abendgymnasien Abendgymnasiums Abendgymnasium
Abendkleide Abendkleides Abendkleids Abendkleider Abendkleid Abendkleidern
Abendkursen Abendkurs Abendkurses Abendkurse
Abendkursen Abendkurse Abendkursus
Abendlands Abendlandes Abendlande Abendland
Abendländers Abendländern Abendländer
Abendmähler Abendmahles Abendmahl Abendmählern Abendmahle Abendmahls
Abendprogramm Abendprogramme Abendprogrammen Abendprogramms
Abendrots Abendrot

Abendröte
Abenteuer Abenteuers Abenteuern
Abenteuerinnen Abenteuerin
Abenteuerlüste Abenteuerlüsten Abenteuerlust
Abenteurer Abenteurern Abenteurers
Abenteurerinnen Abenteurerin
Aberglauben Aberglaube Aberglaubens
Aberglauben Aberglaubens
Aberkennungen Aberkennung
Aberration Aberrationen
Aberwitzes Aberwitzen Aberwitze Aberwitz
Abfahrten Abfahrt
Abfahrtszeit Abfahrtszeiten
Abfalles Abfalls Abfälle Abfällen Abfall Abfalle
Abfalleimern Abfalleimer Abfalleimers
Abfallprodukt Abfallprodukten Abfallprodukte Abfallprodukts Abfallproduktes
Abfassungen Abfassung
Abfertigung Abfertigungen
Abfindungen Abfindung
Abfindungssumme Abfindungssummen
Abflug Abflugs Abflüge Abfluge Abflügen Abfluges
Abflusses Abflüssen Abflüsse Abfluß Abflusse
Abflußhahns Abflußhähnen Abflußhahn Abflußhahnes Abflußhähne Abflußhahne
Abfolge Abfolgen
Abfuhren Abfuhr
Abführmittel Abführmittels Abführmitteln
Abführungen Abführung
Abfütterung Abfütterungen
Abgaben Abgabe
Abgabesoll Abgabesolls
Abgänger Abganges Abgange Abgängern Abgang Abgangs
Abgangszeugnis Abgangszeugnisse Abgangszeugnissen Abgangszeugnisses
Abgasen Abgases Abgase Abgas
Abgefeimtheit Abgefeimtheiten
Abgeklärtheit
Abgeld Abgelde Abgeldes Abgelder Abgelds Abgeldern
Abgeltungen Abgeltung
Abgeordnetem Abgeordneter Abgeordnete Abgeordnetes Abgeordneten
Abgeordnetenhäusern Abgeordnetenhause Abgeordnetenhauses Abgeordnetenhäuser
Abgeordnetenhaus
Abgesandte Abgesandtem Abgesandter Abgesandten Abgesandtes
Abgesangs Abgesang Abgesängen Abgesänge Abgesange Abgesanges
Abgeschiedenheit Abgeschiedenheiten
Abgeschmacktheit Abgeschmacktheiten
Abgestumpftheit
Abgewogenheit
Abglanz Abglanzes Abglanze
Abgottes Abgötter Abgotte Abgotts Abgöttern Abgott
Abgottschlange Abgottschlangen
Abgrenzungen Abgrenzung
Abgrunds Abgründen Abgrundes Abgründe Abgrunde Abgrund
Abguß Abgüssen Abgüsse Abgusse Abgusses
Abgängern Abgängers Abgänger
Abgöttereien Abgötterei
Abhaltung Abhaltungen
Abhandlung Abhandlungen
Abhanges Abhänge Abhange Abhängen Abhangs Abhang
Abhilfe
Abhitze
Abholer Abholern Abholers
Abholzungen Abholzung

```
Abhängigkeit Abhängigkeiten
Abhängigkeitsverh"altnissen Abhängigkeitsverh"altnisses
Abhängigkeitsverh"altnisse Abhängigkeitsverh"altnis
```

5.11 Set oznaka STTS-a

ADJA	adjective, attributive	[das] große [Haus]
ADJD	adjective, adverbial or predicative	[er fährt] schnell, [er ist] schnell
ADV	adverb	schon, bald, doch
APPR	preposition; circumposition left	in [der Stadt], ohne [mich]
APPRART	preposition with article	im [Haus], zur [Sache]
APPO	postposition	[ihm] zufolge, [der Sache] wegen
APZR	circumposition right	[von jetzt] an
ART	definite or indefinite article	der, die, das, ein, eine, ...
CARD	cardinal number	zwei [Männer], [im Jahre] 1994
FM	foreign language material	[Er hat das mit “] A big fish [” übersetzt]
ITJ	interjection	mhm, ach, tja
KOUI	subordinate conjunction with *zu* and infinitive	um [zu leben], anstatt [zu fragen]
KOUS	subordinate conjunction with sentence	weil, daß, damit, wenn, ob
KON	coordinate conjunction	und, oder, aber
KOKOM	comparative conjunction	als, wie
NN	common noun	Tisch, Herr, [das] Reisen
NE	proper noun	Hans, Hamburg, HSV
PDS	substituting demonstrative pronoun	dieser, jener
PDAT	attributive demonstrative pronoun	jener [Mensch]
PIS	substituting indefinite pronoun	keiner, viele, man, niemand
PIAT	attributive indefinite pronoun without determiner	kein [Mensch], irgendein [Glas]
PIDAT	attributive indefinite pronoun with determiner	[ein] wenig [Wasser], [die] beiden [Brüder]
PPER	non-reflexive personal pronoun	ich, er, ihm, mich, dir
PPOSS	substituting possessive pronoun	meins, deiner
PPOSAT	attributive possessive pronoun	mein [Buch], deine [Mutter]

PRELS	substituting relative pronoun	[der Hund ,] der
PRELAT	attributive relative pronoun	[der Mann ,] dessen [Hund]
PRF	reflexive personal pronoun	sich, dich, mir
PWS	substituting interrogative pronoun	wer, was
PWAT	attributive interrogative pronoun	welche [Farbe], wessen [Hut]
PWAV	adverbial interrogative or relative pronoun	warum, wo, wann, worüber, wobei
PAV	pronominal adverb	dafür, dabei, deswegen, trotzdem
PTKZU	*zu* before infinitive	zu [gehen]
PTKNEG	negative particle	nicht
PTKVZ	separable verbal particle	[er kommt] an, [er fährt] rad
PTKANT	answer particle	ja, nein, danke, bitte
PTKA	particle with adjective or adverb	am [schönsten], zu [schnell]
SGML	SGML markup	<turnid=n022k_TS2004>
SPELL	letter sequence	S-C-H-W-E-I-K-L
TRUNC	word remnant	An- [und Abreise]
VVFIN	finite verb, full	[du] gehst, [wir] kommen [an]
VVIMP	imperative, full	komm [!]
VVINF	infinitive, full	gehen, ankommen
VVIZU	Infinitive with *zu*, full	anzukommen, loszulassen
VVPP	perfect participle, full	gegangen, angekommen
VAFIN	finite verb, auxiliary	[du] bist, [wir] werden
VAIMP	imperative, auxiliary	sei [ruhig !]
VAINF	infinitive, auxiliary	werden, sein
VAPP	perfect participle, auxiliary	gewesen
VMFIN	finite verb, modal	dürfen
VMINF	infinitive, modal	wollen
VMPP	perfect participle, modal	gekonnt, [er hat gehen] können
XY	non-word containing non-letter	3:7, H2O, D2XW3
$,	comma	,
$.	sentence-final punctuation mark	. ? ! ; :
$(	other sentence-internal punctuation mark	- [,]()

5.12 Set oznaka TIGER korpusa za obilježivač vrsta riječi

5.12.1 Set oznaka za čvorove

AA	superlative phrase with *am*
AP	adjective phrase
AVP	adverbial phrase
CAC	coordinated adposition
CAP	coordinated adjective phrase
CAVP	coordinated adverbial phrase
CCP	coordinated complementiser
CH	chunk
CNP	coordinated noun phrase
CO	coordination
CPP	coordinated adpositional phrase
CS	coordinated sentence
CVP	coordinated verb phrase (non-finite)
CVZ	coordinated infinitive with *zu*
DL	discourse level constituent
ISU	idiosyncratic unit
MTA	multi-token adjective
NM	multi-token number
NP	noun phrase
PN	proper noun
PP	adpositional phrase
QL	quasi-language
S	sentence
VP	verb phrase (non-finite)
VZ	infinitive with *zu*

5.12.2 Set rubnih oznaka TIGER korpusa

AC	adpositional case marker
ADC	adjective component
AG	genitive attribute
AMS	measure argument of adjective
APP	apposition
AVC	adverbial phrase component
CC	comparative complement
CD	coordinating conjunction
CJ	conjunct
CM	comparative conjunction
CP	complementizer
CVC	collocational verb construction (*Funktionsverbgefüge*)
DA	dative
DH	discourse-level head
DM	discourse marker
EP	expletive *es*
HD	head
JU	junctor
MNR	postnominal modifier
MO	modifier
NG	negation
NK	noun kernel element
NMC	numerical component
OA	accusative object
OA	second accusative object
OC	clausal object
OG	genitive object
OP	prepositional object
PAR	parenthetical element
PD	predicate
PG	phrasal genitive
PH	placeholder
PM	morphological particle
PNC	proper noun component
RC	relative clause
RE	repeated element
RS	reported speech
SB	subject
SBP	passivised subject (PP)
SP	subject or predicate
SVP	separable verb prefix
UC	unit component
VO	vocative

6 Izvori

Silić, J., Pranjković, I. (2005). *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*. Školska Knjiga, Zagreb

Jurafsky, D., Martin, J. H. (2000). *Speech and language Processing: An introduction to natural Language processing, Computational Linguistics and Speech Recognition*. Prentice Hall

G.J. Babu (2012). *Bayesian and frequentist approaches*. Online Proceedings of the Astronomical Data Analysis Conference (ADA VII)

Hentschel, E., Vogel, P. M. (2009). *Deutsche Morphologie*. de Gruyter, Berlin/New York

Dudenredaktion (2011). *Duden, Deutsches Universalwörterbuch, 7., übrearbeitete und erweiterte Auflage*. Dudenverlag. Manheim, Zürich

Bußmann, H. (2002). *Lexikon der Sprachwissenschaft*. 3., aktualisierte und erweiterte Auflage. Kröner, Stuttgart

Savoy, J. (2000). *Light stemming approaches for the French, Portuguese, German and Hungarian languages*. Institut interfacultaire d'informatique, University of Neuchate. Neuchâtel, Switzerland

Weissweiler, L., Fraser, A. (2017). *Developing a Stemmer for German Based on a Comparative Analysis of Publicly Available Stemmers*. Center for Information and Language Processing, LMU Munich, Germany

Manning, C. D., Raghavan, P., Schütze, H (2009). *An Introduction to Information Retrieval*. Cambridge University Press Cambridge, England

Sidorov, G. (2013). *Syntactic Dependency Based N-grams in Rule Based Automatic English as Second Language Grammar Correction*. International Journal of Computational Linguistics and Applications S 168-188

Pobar, M., Martinčić-Ipšić, S., Ipšić, I. (2008). *Računalni sustav za tvorbu hrvatskoga govora text-to-speech synthesis: a prototype system for croatian language*.

Lioma, C., van Rijsbergen, C.J. K. (2008). *Part of speech n-grams and Information Retrieval*. Revue française de linguistique appliquée S. 9-22

Hecht, R., Riedler, J, Backfried, G. (2002). *Fitting German into N-Gram Language Models*. Speech, Artificial Intelligence, and Language Laboratories, Vienna, Austria

Wang, T. (2015). *Hidden Markov Model Based Recognition of German Finger Spelling Using the Leap Motion*. Faculty of Natural Sciences and Technology I Department of Computer and Communication Technology, Saarland University

Gagniuc, P. A. (2017). *Markov Chains: From Theory to Implementation and Experimentation*.: John Wiley & Sons. USA, New Jersey S 1–256.

Weißweiler, L., Fraser, A. (2017). *Developing a Stemmer for German Based on a Comparative Analysis of Publicly Available Stemmers*. Center for Infformation and Language Processing, LMU Munich

Caumanns, J. (1999). *A Fast and Simple Stemming Algorithm for German Words*. Technical Report Nr. trb-99-16. Freie Universitat Berlin, Fachbereich Math- ematik und Informatik

Manning, C. D., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge

Porter, M. F. (1980). *An algorithm for suffix stripping*. Program, 14(3):130–137.

Moortgat, M. (2001). *Logical Aspects of Computational Linguistics*. Die Deutsche Bibliothek, Berlin

Sennrich, R., Kunz, B. (2014). *Zmorge: A German Morphological Lexicon Extracted from Wiktionary*. Institute of Computational Linguistics, University of Zürich

Adolphs, P. (2008). *Acquiring a Poor Man's Inflectional Lexicon for German*. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, Marrakech, Morocco

Haapalainen, M., Majorin, A. (1995). *GERTWOL und Morphologische Disambiguierung für das Deutsche*. In Proceedings of the 10th Nordic Conference of Computational Linguistics, Helsinki

Schmid, H., Fitschen, A., Heid, U. (2004). *SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection*. Proceedings of the IVth International Conference on Language Resources and Evaluation, LREC S:1236-1266, Lisabon, Portugal
Schütze, H., Singer, Y. (1994). *Part-of-speech tagging using a variable memory Markov model*. 32nd Annual Meeting of the Association for Computational Linguistics. S. 181-187
Telljohann, H., Hinrichs, E., Kübler, S. (2004). *The Tüba-D/Z Treebank: Annoting German with a Context-Free Backbone*. Seminar für Sprachwissenschaft. Universität Tübingen, Deutschland
Rhorer, C., Forst, M. : „*Improving coverage and parsing qualityof a large-scale LFG for German*“. Stuttgart. Njemačka

Kübler, S. (2008.): „*The PaGe 2008 shared task on parsing German* “. Indiana University. Bloomington. SAD

Neumann, D., Braun, C., Piskorski, J. (2012.): “*A Devide-And-Conquer strategy for shallow Parsing of German*”. University of Zürich. Switzerland

Smith, G. (2003): „*A brief introduction to the TIGER treebank, Version 1*“. Universität Potsdam. Germany

Crysmann, B., Hansen-Schirra, S., Smith, G., Zigler-Eisele, D. (2005): „*TIGER Morphologie-Annotationsschema*“. Universität Stuttgart. Germany

Brants, S., Dipper, S., Hansen, S., Lezius, W., Smith, G. (2002): „*The TIGER Treebank* “. Saarland Univerity. Germany

Rehbein, I., van Genabith, J. (2007). *Trebnak annotiation Svhemes and parser evolution for German*. Association for Computational Linguistics. Chech Republic

Schiehlen, M. (2008): “*Annotation Strategies for Porbabilistic parsing in German*”. Univerisity of Stuttgart. Germany

Sennrich, R., Schneider, G., Volk, M., Warin, M. (2014). *A new hybrid parser for German*. Universität Zürich. Schweiz

Smith, G. (2003): *Searching for Morphological Structure with Regular Expressions*. Universität Potsdam. Germany

Smith, G. (2003). *A Brief Introduction to the TIGER Treebank, Version 1*. Universität Potsdam

Feldweg, H. (1999). *Implementation and evaluation of a german HMM for POS disambiguation*. Springer, Dordrecht

Sennrich, R:, Schneider, G., Volk, M., Warin, M. (2009). *A New Hybrid Dependency Parser for German*. Universität Zürich, Institut für Computerlinguistik, Zürich

Jurish, B. (2003*). A Hybrid Approach to Part-of-Speech Tagging*. Berlin-Brandenbursche Akademie der Wissenschaften

Markov Analysis by Dr. V.V. HaraGopal- SlideShare (12.5.2018). https://www.slideshare.net/ganith2k13/markov-analysis

Markov chain-Oxford Dictionaries (12.5.2018). https://en.oxforddictionaries.com/definition/us/markov_chain

Hrvatska Enciklopedija-Semantika (8.5.2018). http://www.enciklopedija.hr/natuknica.aspx?id=49926

Hrvatska Enciklopedija-Fonetika (8.5.2018). http://www.enciklopedija.hr/Natuknica.aspx?ID=20060

Hrvatska Enciklopedija-Fonologija (8.5.2018). http://www.enciklopedija.hr/natuknica.aspx?id=20069

SAS-Natural Language Processing. What it is and why it matters (8.5.2018). https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

Snowball Tartarus- German stemming algorithm (8.5.2018). http://snowball.tartarus.org/algorithms/german/stemmer.html

German Porter Stemmer in JavaScript-GitHub (12.5.2018). https://gist.github.com/marians/942312

Gemranic language stemmer-Snowball.Tartarus (17.5.2018). http://snowball.tartarus.org/texts/germanic.html

IWNLP.Lemmatizer-GitHub (16.5.2018). https://github.com/Liebeck/IWNLP.Lemmatizer

Zeit.de (22.08.2018.). https://www.zeit.de/politik/ausland/2017-06/malta-joseph-muscat-labour-neuwahlen-panama-papers